# Ipseology

**A new science of the self**

Dr. Jason Jeffrey Jones

2023-11-27

# Table of contents

# Preface

Ipseology is the study of human identity using large datasets and computational methods.

This book introduces the main ideas and assumes no prior experience. It is written by me, Dr. Jason Jeffrey Jones.

Your feedback is welcome. For now, read on!

Go to Chapter 1

## Citation

Jones, J. J. (2023). *Ipseology - A new science of the self.* Jason Jeffrey Jones Productions. ISBN 979-8-9896544-0-6 https://jasonjones.ninja/ipseology-a-new-science-of-the-self-book/

This book is a living document, begun in December 2022 and first "completed" on November 27, 2023. I added material and reorganized on March 15, 2024. New chapters may appear in the future. The most up-to-date and definitive version lives on the web at https://jasonjones.ninja/ipseology-a-new-science-of-the-self-book/

# 1 Introduction

## 1.1 What is ipseology?

Ipseology is a word I made up. It refers to the study of ipseity. My definition of ipseity is: *selfhood, individuality and the elements of identity.*

I want to study ipseity **at scale**. At scale means with millions of observations covering substantial temporal periods and geographical areas. Think 20 million American users of Twitter over six years to start. Then ask, why not hundreds of millions across dozens of nations over more than a decade?

Let's define ipseology as the study of human identity using large datasets and computational social science methods.

## 1.2 How does one do ipseology research?

Be patient; that's what this whole book will tell you!

But for a first peek at the potential, let's look at a fast-growing trend in Twitter profile biographies. Ipseology is ideal for studying temporal trends (i.e. changes in the language individuals use to describe themselves). Look at Figure 1.1.

In 2017, very few users included pronoun lists such as `she/her`, `he/him` and `they/them` in their profile biographies. Over the next few years, the numbers increased rapidly:

I formally define *prevalence* in Chapter 3.

- From January 1, 2017 to June 30, 2022, the prevalence of `she/her` increased nearly 20x from 16 to 313.
- The prevalence of `he/him` increased nearly 28x from 10 to 276.
- The prevalence of `they/them` increased 5x from 11 to 58.

The ability to perform analysis like this is exciting for many reasons.

First, these estimates are more **precise** than is typical in social science analysis. Typical social science estimates are made based on small samples - from N=1 at worst to N=1,000 for a well-funded study. Each estimate in the above is drawn from about 200,000 observations.

Second, these estimates have finer **resolution** than is typical in social science analysis. In typical social science methods, the researcher measures exactly once, maybe twice if funding allows. With the ipseological approach, one instead estimates each and every day based on thousands of new observations.

Precision and resolution are just two advantages of the ipseological approach. For now, suffice to say, we have a new way to examine human identities. We can use it consistently across time and geography. Let's get started! Chapter 2

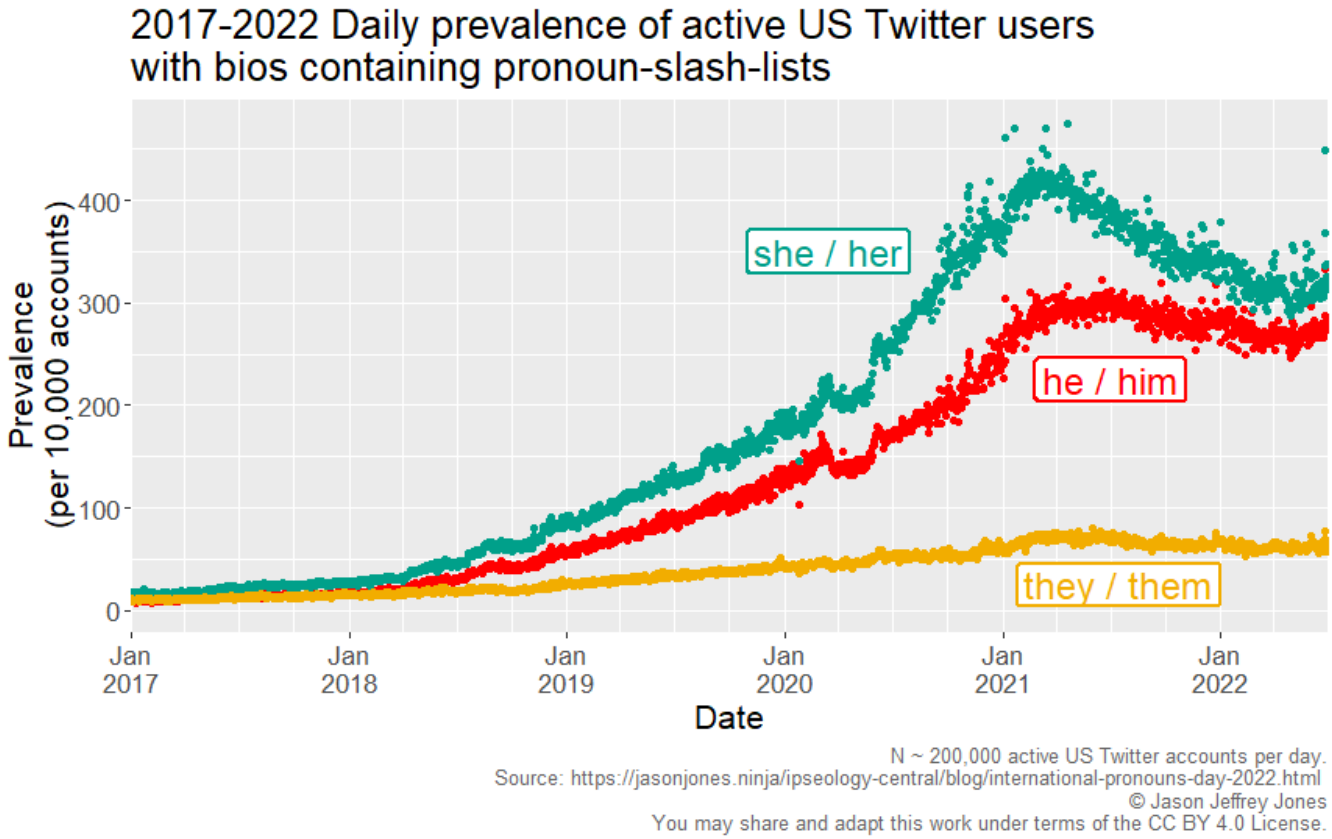


Figure 1.1: Daily prevalence of active US Twitter users with bios containing pronoun lists

# 2 How to think like an ipseologist

## 2.1 Personally expressed identity is the data.

One measures ipseity with **language**. Specifically, the data sought is personally expressed identity text. This text must satisfy all three conditions:

- It should be personal - the authors are describing themselves.
- It should be expressed - the authors' text is "published," i.e. the words are available where others might see them.
- It should describe identity - the explicit purpose of the text is description of the author.

Currently, personally expressed identity text data is best sourced from social media profiles. For the years 2012-2023, Twitter profile biographies (bios) were the best source because of the scale at which the platform was used and the open availability of public profile data.

## 2.2 Individuals are bags of signifiers.

Individuals use words to describe themselves. Ipseologists presume these words are a strong signal of self-perceived identity. Others worry whether these words should be considered veridical, aspirational, performative or some other category. Ipseologists don't worry. Or more accurately, we don't pretend that there is some gauge of the 'true self' that a bio fails to comport with.

When we want to know who someone is, we ask them. Then we analyze the language.

Specifically, we look at the signifiers (words) they choose. Consider @BarackObama's bio: `Dad, husband, President, citizen`. Here is a model subject - no function words, just a comma-delimited list of identity signifiers.

Everyone is a bag of signifiers. Presuming this vastly simplifies analysis. Each individual, at each time point is a small set of self-ascribed words. Analysis can begin with mere counting. How many individuals consider themselves `dad` and how many `mom`? Small increases in analysis complexity reap big rewards. One can proceed to measure which signifiers co-occur with `mom` but not `dad`. If an individual removes the word `dad` from his bio, which signifier is he most likely to replace it with?

We will look into these more sophisticated questions soon, but let's first build a foundation with simpler analyses. Chapter 3

# 3 Understanding ipseological analysis

## 3.1 Counting is a good place to begin.

How many users include the word `vegan` in their Twitter profile? I can tell you the exact number I have observed for one place and one time. The number is 7,316 for American users in 2012. Call that number the ***incidence*** - it's a raw count.

Now say I want to compare 2012 to 2020. In 2020, the incidence of US `vegan` users was 13,756. Wow, that's a lot more! But you suspect maybe there were a different number of Twitter users to be observed in 2020 than in 2012, and you are correct. The total number of unique US accounts I observed in 2012 was about 10 million and in 2020 there were about 10.2 million, i.e. 200,000+ more opportunities to observe `vegan`s.

I have included code blocks for those who wish to see implementation details.

```
library(tidyverse)

# Download a csv file containing US data for tokens at annual resolution.
# Read about the data in the text file at https://osf.io/3gjta
tac = read_csv("https://osf.io/download/cdzsb/")

# Become familiar with the data file.
str(tac) # Structure of the data.
tac %>% slice_sample(n = 5)  # Example rows.

# How many users include the word vegan in their Twitter profile?
tac %>% filter(token == "vegan" & obsYear == 2012) # in 2012
tac %>% filter(token == "vegan" & obsYear == 2020) # in 2020
# Note that the numerator column contains the incidence.
```

## 3.2 Prevalence affords comparison.

We need to do some simple math to compare apples-to-apples. We *might* divide the incidence by the total number of observed accounts to get a proportion. When we do so, we find `vegan` proportions of 0.000735 for 2012 versus 0.001351 in 2020.

```
# Calculate ugly proportions.
tac %>% filter(token == "vegan" & (obsYear == 2012 | obsYear == 2020)  ) %>%
  mutate(proportion = numerator / denominator)

# Calculate pretty prevalances.
tac %>% filter(token == "vegan" & (obsYear == 2012 | obsYear == 2020)  ) %>%
  mutate(prettyPrevalence = round(10000* numerator / denominator) )
```

**Yuck.** People - even smart ones like you - have trouble with fractional numbers like these. When asked to compare fractions, they take longer and still make more mistakes then they do when comparing counting numbers. To avoid this unnecessary cognitive hurdle, I consistently express the popularity of signifiers in bios in terms of prevalence per 10,000. Consider Figure 3.1. It shows prevalence for `vegan` among US users for the years 2012-2023. The y-axis has whole number, human-friendly units.

```
# Visualize vegan prevalence over time.
tac %>% filter(token == "vegan") %>%
  mutate(finePrevalence = 10000 * numerator / denominator ) %>%
  mutate(Signifier = factor(token, levels = c("vegan")) ) %>%
ggplot(aes(x = obsYear, y = finePrevalence, color = Signifier, shape = Signifier)) +
  geom_path(linetype = "longdash", linewidth = 1) +
  geom_point(size=4) +
  scale_x_continuous(limits=c(2012,2023), breaks = 2012:2023, minor_breaks = NULL) +
  scale_color_manual(values = c("#5c8326")) +
  ggtitle("Estimated prevalence of vegan signifier", "within US Twitter users' profile bios 2012-2023
  xlab("Year") + ylab("Prevalence\n(per 10,000 accounts)") +
  theme(text = element_text(size=14)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(legend.position = c(0.75, 0.5),
      legend.background = element_rect(fill = "white", color = "black")) +
  labs(caption = "Source: Ipseology - a new science of the self\n \uA9 Jason Jeffrey Jones. You may s
  theme(plot.caption = element_text(size=10, color = "#666666"))
```

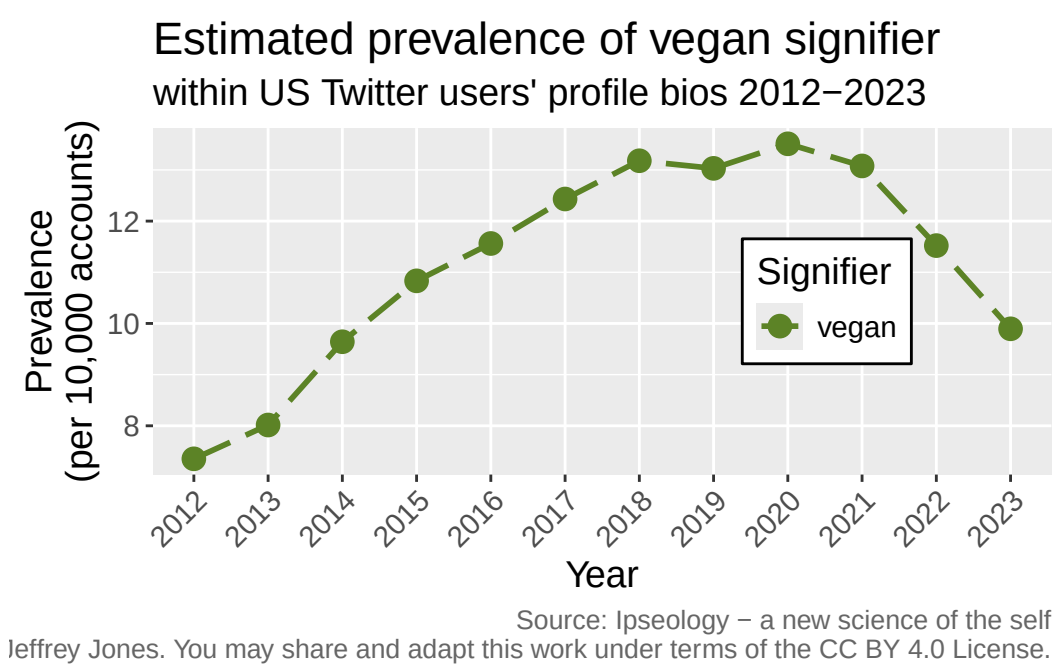


Figure 3.1: Prevalence of vegan over time.

Now we have a consistent, simple unit of measurement. ***Prevalence*** allows comparison across time *and* between signifiers. Consider Figure 3.2. Among the observed users, clearly `vegan` increased in popularity while `vegetarian` decreased. `Carnivore` rose above threshold in 2019 and grew through 2023.

(The threshold to be included in the data was 1 or more per 10,000 after rounding. `omnivore` and `pescatarian` never exceeded threshold in any of these years. Using a threshold eliminates the long tail of signifiers used by only tiny fractions of the user population and all tokens that would uniquely identify one person.)

```
# Multiple signifier series.
tac %>% filter(token %in% c("vegan", "vegetarian", "carnivore") ) %>%
  mutate(finePrevalence = 10000 * numerator / denominator ) %>%
  mutate(Signifier = factor(token, levels = c("vegan", "vegetarian", "carnivore")) ) %>%
ggplot(aes(x = obsYear, y = finePrevalence, color = Signifier, shape = Signifier)) +
  geom_path(linetype = "longdash", linewidth = 1) +
  geom_point(size=4) +
  scale_x_continuous(limits=c(2012,2023), breaks = 2012:2023, minor_breaks = NULL) +
  scale_color_manual(values = c("#5c8326", "#B61500", "#a16868") ) +
  ggtitle("Estimated prevalence of food-related signifiers", "within US Twitter users' profile bios 2
  xlab("Year") + ylab("Prevalence\n(per 10,000 accounts)") +
  theme(text = element_text(size=14)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(legend.position = c(0.75, 0.55),
      legend.background = element_rect(fill = "white", color = "black")) +
  labs(caption = "Source: Ipseology - a new science of the self\n \uA9 Jason Jeffrey Jones. You may s
  theme(plot.caption = element_text(size=10, color = "#666666"))
```

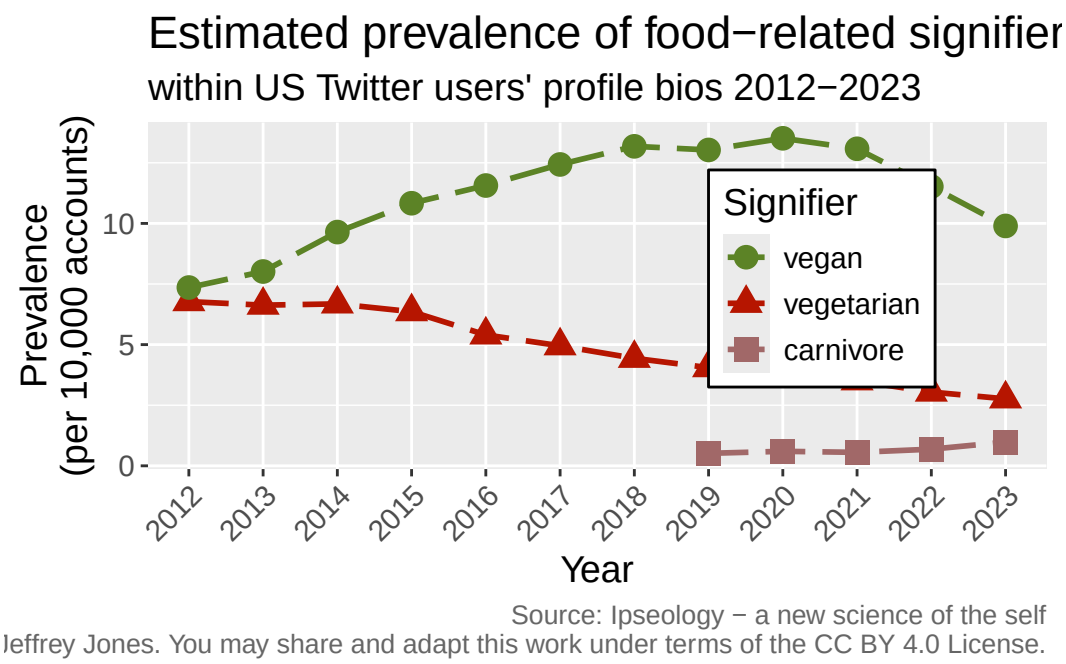


Figure 3.2: Prevalence of vegan, vegetarian and carnivore over time. Note that we have measured the popularity of signifiers consistently, persistently and precisely.

Using the prevalence measure, we can perform analysis over time. What about space? Fortunately, our Twitter data is multinational; let's compare prevalence of the same signifiers across countries. Chapter 4

## 3.3 Data and code.

Download data for this chapter from https://osf.io/download/cdzsb/

R code to produce the numbers and figures within this chapter is embedded in the code chunks above.

# 4 Let's take ipseology global!

## 4.1 A hesitant peek outside the United States.

So far, we have only analyzed data from users in the United States. The Twitter data was sampled from everywhere, however. Why not extend our analyses to more nations?

We can begin with similar places. Australia, Canada and the United Kingdom share the primary language of English with the US. I would expect to see most trends observed in the US also present in these countries because of the strong web of influence the cultures are enmeshed in. Let's recreate Figure 3.2 for each of these nations.

```
library(tidyverse)

# Download a csv file containing multinational data for tokens at annual resolution.
# Read about the data in the text file at https://osf.io/mdp7k
# HINENI stands for Human Identity across Nations of the Earth Ngram Investigator.
hineni = read_csv("https://osf.io/download/k7bwj/")

# Become familiar with the data file.
str(hineni) # Structure of the data.
hineni %>% slice_sample(n = 5)  # Example rows.

# How many users include the word vegan in their Twitter profile?
hineni %>% filter(ngram == "vegan" & obsYear == 2012) # in 2012
hineni %>% filter(ngram == "vegan" & obsYear == 2020) # in 2020
# Note that the numerator column contains the incidence.
```

```
hineni %>%
  filter(ngram %in% c("vegan", "vegetarian", "carnivore") ) %>%
  filter(nation %in% c("AU", "CA", "GB", "US") ) %>%
  mutate(Nation = factor(nation, levels = c("AU", "CA", "GB", "US"), labels = c("Australia", "Canada'
  mutate(finePrevalence = 10000 * numerator / denominator ) %>%
  mutate(Signifier = factor(ngram, levels = c("vegan", "vegetarian", "carnivore")) ) %>%
ggplot(aes(x = obsYear, y = finePrevalence, color = Signifier, shape = Signifier)) +
  geom_path(linetype = "longdash", linewidth = 1) +
  geom_point(size=4) +
  scale_x_continuous(breaks = seq(2012, 2023, 2)) +
  scale_color_manual(values = c("#5c8326", "#B61500", "#a16868") ) +
  ggtitle("Estimated prevalence of food-related signifiers", "within AU, CA, UK and US Twitter users'
  xlab("Year") + ylab("Prevalence\n(per 10,000 accounts)") +
  facet_wrap(vars(Nation), nrow = 2) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1) ) +
  theme(text = element_text(size=16)) +
  theme(legend.background = element_rect(fill = "white", color = "black")) +
  labs(caption = "Source: Ipseology - a new science of the self\n \uA9 Jason Jeffrey Jones.\nYou may
  theme(plot.caption = element_text(size=10, color = "#666666"))
```

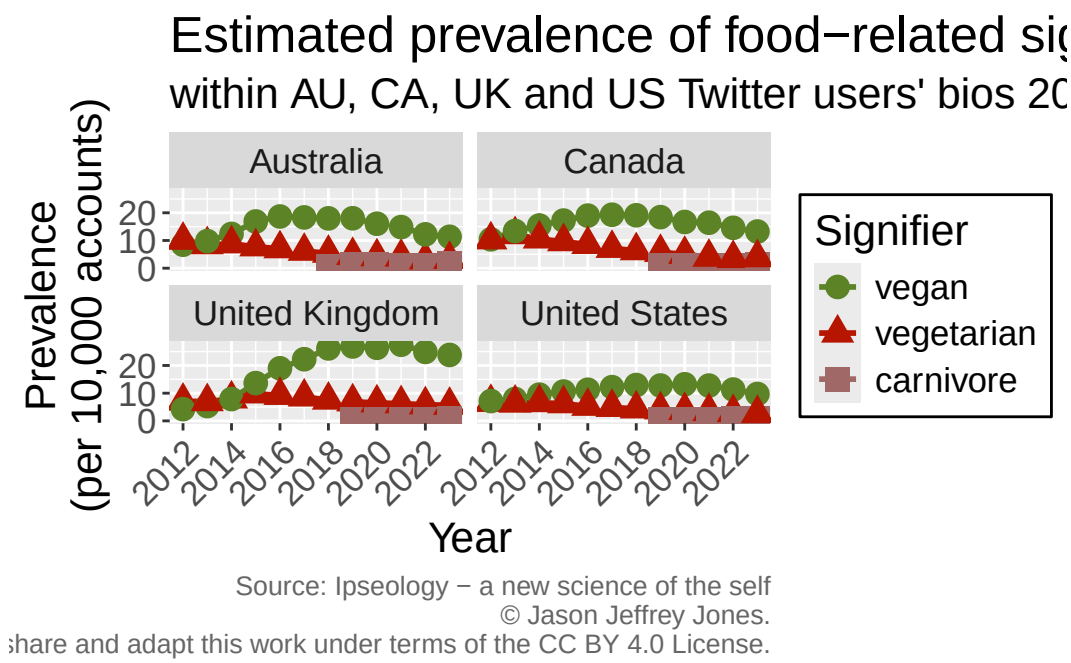


Figure 4.1: Prevalence of vegan, vegetarian and carnivore over time in Australia, Canada, the United Kingdom and the United States.

We see the same pattern in each country: `vegan` growth and `vegetarian` decline. Compared to the United States, the rise of `vegan` was more vigorous in Australia, Canada and especially the United Kingdom.

## 4.2 Use previous analysis as a template.

We begin to see the utility of the ipseological approach. When we measure consistently, persistently and precisely at scale, comparisons across time and space become accessible.

Let's use our previous analysis as a template for more. There are several primarily Spanish-speaking nations in the data. Why not try a simple extension? Google Translate tells me there are feminine and masculine forms of vegan (vegana, vegano) and similarly for vegetarian (vegetariana, vegetariano). How popular was each of these across years and nations?

```
hineni %>%
  filter(ngram %in% c("vegano", "vegana", "vegetariano", "vegetariana") ) %>%
  filter(nation %in% c("AR", "CL", "CO", "MX", "PE", "ES", "VE") ) %>%
  mutate(Gender = if_else(ngram %in% c("vegana", "vegetariana"), "Fem", "Masc") ) %>%
  mutate(Nation = factor(nation, levels = c("AR", "CL", "CO", "MX", "PE", "ES", "VE"), labels = c("Ar
  mutate(finePrevalence = 10000 * numerator / denominator ) %>%
  mutate(Signifier = factor(ngram, levels = c("vegano", "vegana", "vegetariano", "vegetariana")) ) %>
  # Get rid of Chile.
  #filter(Nation != "Chile" ) %>%
ggplot(aes(x = obsYear, y = finePrevalence, color = Signifier, shape = Signifier)) +
  geom_path(linetype = "dashed", linewidth = 0.75) +
  geom_point(size=2) +
  scale_x_continuous(breaks = seq(2012, 2023, 2)) +
  #scale_y_continuous(limits = c(0,6), breaks = seq(0, 6, 2), expand =  expansion(add=c(0.5, 2)) ) +
  scale_color_manual(values = c("#5c8326", "#5c8326", "#B61500", "#B61500") ) +
  ggtitle("Estimated prevalence of food-related signifiers", "within some Spanish-speaking nations' T
```

```
  xlab("Year") + ylab("Prevalence\n(per 10,000 accounts)") +
  facet_grid(rows = vars(Nation), cols = vars(Gender) ) +
  # , scales = "free_y"
  theme(axis.text.x = element_text(angle = 45, hjust = 1) ) +
  #theme(text = element_text(size=16)) +
  theme(strip.text.y = element_text(angle = 0) ) +
  #theme(panel.spacing.y = unit(2, "lines") ) +
  theme(legend.background = element_rect(fill = "white", color = "black")) +
  labs(caption = "Source: Ipseology - a new science of the self\n \uA9 Jason Jeffrey Jones.\nYou may
  theme(plot.caption = element_text(size=10, color = "#666666"))
```

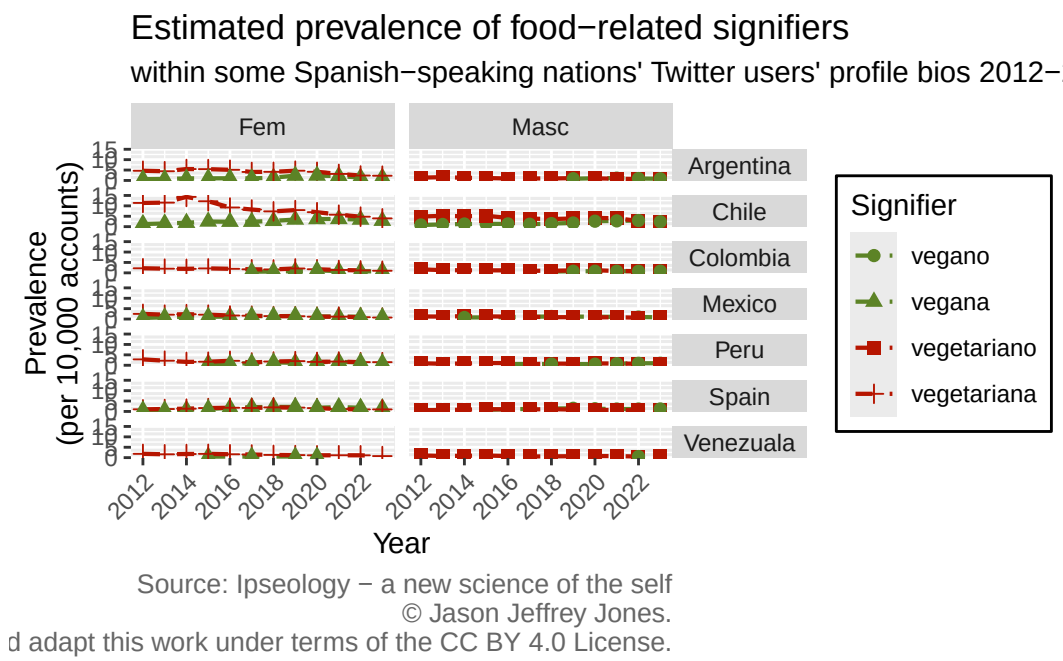


Figure 4.2: Prevalence of feminine and masculine forms of vegan and vegetarian over time in some Spanish-speaking nations.

This is not the most beautiful figure, but it tells us quite a bit. Our chosen signifiers - `vegano`, `vegana`, `vegetariano`, `vegetariana` saw infrequent use. Prevalence was in the low single digits with the exceptions of early years `vegetariana` in Chile.

Some (not many) define themselves by what they eat. How about what one plays or watches?

## 4.3 The multinational ipseology of soccer.

Let's stick with Spanish-speaking nations and examine two ways individuals might signify their interest in what Pelé called the beautiful game: `fútbol` and the `:soccer:` emoji.

```
# Load emojis from file.
library(emoji)

soccerTokens = c("fútbol", emoji('soccer'))

hineni %>%
```

```
  filter(ngram %in% soccerTokens ) %>%
  filter(nation %in% c("AR", "CL", "CO", "MX", "PE", "ES", "VE") ) %>%
  mutate(Nation = factor(nation, levels = c("AR", "CL", "CO", "MX", "PE", "ES", "VE"), labels = c("Ar
  mutate(finePrevalence = 10000 * numerator / denominator ) %>%
  mutate(Signifier = factor(ngram, levels = soccerTokens) ) %>%
ggplot(aes(x = obsYear, y = finePrevalence, color = Signifier, shape = Signifier)) +
  geom_path(linetype = "dashed", linewidth = 0.75) +
  geom_point(size=2) +
  scale_x_continuous(breaks = seq(2012, 2023, 2)) +
  #scale_y_continuous(limits = c(0,6), breaks = seq(0, 6, 2), expand =  expansion(add=c(0.5, 2)) ) +
  #scale_color_manual(values = c("#5c8326", "#5c8326", "#B61500", "#B61500") ) +
  ggtitle("Estimated prevalence of fútbol signifiers", "within some Spanish-speaking nations' Twitter
  xlab("Year") + ylab("Prevalence\n(per 10,000 accounts)") +
  facet_wrap(vars(Nation), nrow = 3) +
  # , scales = "free_y"
  theme(axis.text.x = element_text(angle = 45, hjust = 1) ) +
  #theme(text = element_text(size=16)) +
  theme(strip.text.y = element_text(angle = 0) ) +
  #theme(panel.spacing.y = unit(2, "lines") ) +
  theme(legend.background = element_rect(fill = "white", color = "black")) +
  labs(caption = "Source: Ipseology - a new science of the self\n \uA9 Jason Jeffrey Jones.\nYou may
  theme(plot.caption = element_text(size=10, color = "#666666"))
```

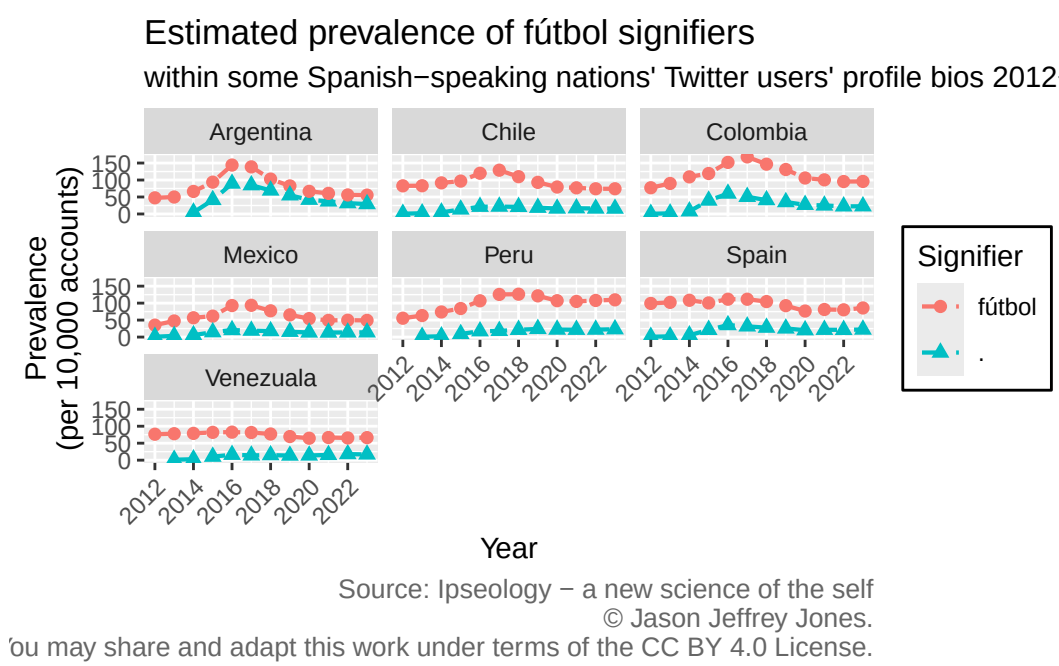


Figure 4.3: Prevalence of soccer-related signifiers in some Spanish-speaking nations.

Note that the maximum on the y-axis for `fútbol` is 10x what it was for `vegetariana`. As ipseologists, we can interpret this difference. Individuals in these countries are vastly more likely to mention a soccer interest as compared to a vegetarian status.

Individuals with `fútbol` in the bio outnumber those with the `:soccer:` emoji in every country and every year. But emojis offer something very compelling. They are language-independent. An arduous way to continue our multinational examination of soccer would be to translate into each language: `Fußball`, `voetbal`, , etc. Let's take a different tack, and follow the wind of emoji instead.

## 4.4 Emojis are language-independent but not context-independent.

Sometimes language differences are exactly what we are interested in; sometimes they just get in the way. Thankfully, there are emojis.

Emojis are language-independent but not context-independent. They are meant to represent the same thing no matter the language or nation they are embedded in. But certainly, some emojis are more personally meaningful to individuals in some contexts compared to others.

In the previous section, we saw some evidence that many individuals in Spanish-speaking nations mention `fútbol` in their self-descriptions. How common is a soccer affinity generally over other nations? How about other sports? Let's use emoji representations of sports to find out.

We have the luxury of data for 32 nations, but let's *not* generate 32 different graphs times 12 years. Instead, let's choose one year (2022) and a reasonable set of sports emojis: , , , , , , .

Now let's visualize the relative popularity of each sport emoji within bios in all these nations.

```
sportsEmojis = c(emoji('soccer'), emoji('basketball'), emoji('baseball'), emoji('cricket game'), emoj
#sportsEmojisColors = c(emoji('soccer') = "#333333", emoji('basketball') = "#ee6730", emoji('baseball
sportsEmojisColors2 = c("#333333", "#ee6730", "#000089",  "#d6b268",  "#663831",  "#0c8c3e",  "#cd346
sportsEmojisColors2 = setNames(sportsEmojisColors2, sportsEmojis)

# Download a file mapping nation abbreviations to nation names.
nationList = read_csv("https://osf.io/download/qat5c/")
nationList = nationList %>% mutate(nationName = factor(nationName) )

hineni %>%
  filter(obsYear == 2022) %>%
  filter(ngram %in% sportsEmojis ) %>%
  mutate(Signifier = factor(ngram, levels = sportsEmojis) ) %>%
  # Join in nationList.
  mutate(nationCode = nation ) %>%
  inner_join(nationList) %>%
  group_by(nationName) %>%
  mutate(proportion = numerator / sum(numerator)) %>%
  ungroup() %>%
  # Order the nationName factor alphabetically.
  mutate(nationName = factor(nationName, levels = sort(unique(nationName), decreasing = TRUE)) ) %>%
  ggplot(aes(x = proportion, y = nationName, fill = Signifier)) +
    geom_bar(stat = "identity") +
    scale_fill_manual(values = sportsEmojisColors2 ) +
    scale_x_continuous(expand = c(0, 0)) +
    ggtitle("Proportion of sports emojis per sport per nation",
            "within each nation's Twitter users' profile bios in 2022") +
    xlab("Proportion") + ylab(NULL) +
    labs(caption = "Source: Ipseology - a new science of the self\n \uA9 Jason Jeffrey Jones.\nYou ma
    theme(plot.caption = element_text(size=10, color = "#666666"))
```

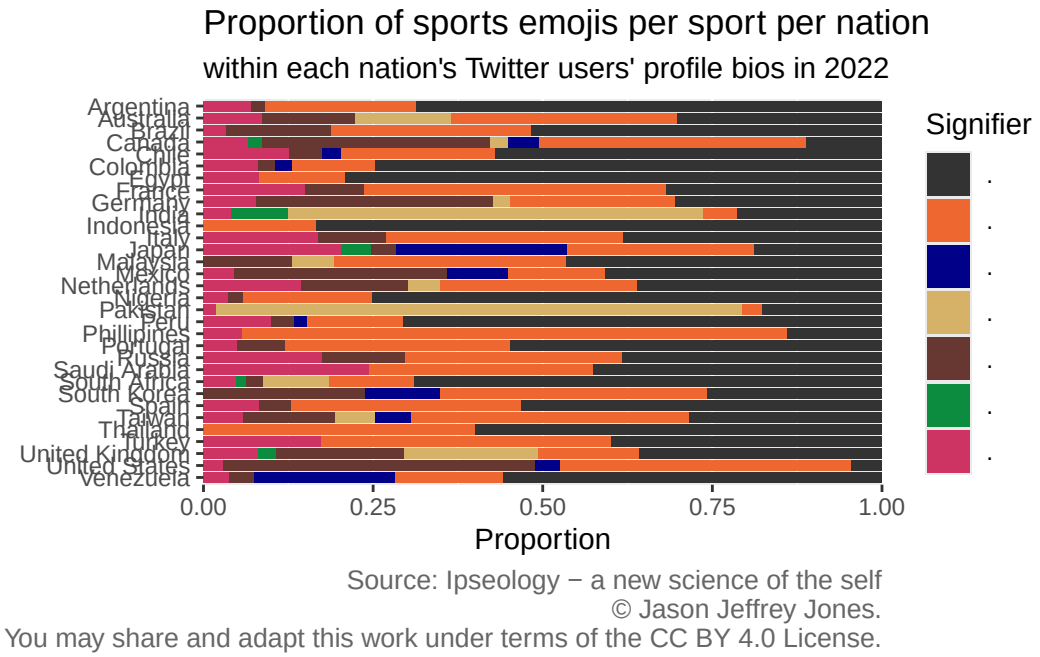


Figure 4.4: Relative popularity of each sport among users in each nation that include a sports emoji in their bio. Find a nation name on the left, and scan across to see the proportions for each sport. Cricket is popular in India, Pakistan and a few other nations. Soccer is widely popular, except in the US, where it is crowded out by basketball and American football.

**The people of earth sure love soccer.** In more than half of these nations,  is the most popular sports emoji users place within their bios.

There are so many more emojis to explore. It just so happened I was thinking about soccer and football when I wrote this chapter. Already, I hope you have thought of more interesting comparisons to make across time and geography. I have tried to make that an easy project to get started on, as you'll see in the next section introducing HINENI.

## 4.5 Explore with HINENI.

HINENI stands for Human Identities across Nations of the Earth Ngram Investigator. HINENI comprises a dataset and tools allowing anyone to explore the popularity of signifiers within profile bios. It covers 32 nations and the years 2012 through 2023.

To explore in a web browser with no coding or download necessary, use the interface Human Identities across Nations of the Earth Ngram Investigator.

To read a peer-reviewed, open access research article, download our ICWSM 2024 article describing and exploring the HINENI dataset.

## 4.6 Data and code.

Download data for this chapter from https://osf.io/download/k7bwj/

R code to produce the numbers and figures within this chapter is embedded in the code chunks above.

# 5 Deleting God, Adding Jesus: Bio Revision Events

## 5.1 What is a *bio revision event*?

In ipseology, a ***bio revision event*** occurs when one observes a change over time in the set of signifiers an individual chooses to include in their self-description. In this Chapter, we'll come to understand the four possible bio revision events: *Add*, *Delete*, *Keep* and *Ignore*.

Let's use religious signifiers in the United States as our domain of study. There are *many* signifiers we might use; let's begin with only two for simplicity: `god` and `jesus`. As Figure 5.1 illustrates, these were consistently popular signifiers within US users' bios. (In 2022, `god` was the 86th most popular token and more popular than 99.4% of all other tokens; `jesus` was more popular than 97.9% of all other tokens.)

```
library(tidyverse)

# Download a csv file containing US data for tokens at annual resolution.
# Read about the data in the text file at https://osf.io/3gjta
tac = read_csv("https://osf.io/download/cdzsb/")

# Compute finePrevalence and begin years at 2012.
tac= tac %>% filter(obsYear > 2011) %>%
  mutate(finePrevalence = 10000 * numerator / denominator )

# For the text, find rankings.
godJesus2022Ranks = tac %>% filter(obsYear == 2022) %>%
  mutate(rowNum = row_number(-finePrevalence) ) %>%
  mutate(minRank = min_rank(-finePrevalence) ) %>%
  mutate(denseRank = dense_rank(-finePrevalence) ) %>%
  mutate(percentRank = percent_rank(finePrevalence) ) %>%
  mutate(cumeDist = cume_dist(finePrevalence) ) %>%
  filter(token %in% c("god", "jesus") )

# Visualize god and jesus prevalence over time.
tac %>% filter(token %in% c("god", "jesus") ) %>%
  mutate(Signifier = factor(token, levels = c("god", "jesus")) ) %>%
ggplot(aes(x = obsYear, y = finePrevalence, color = Signifier, shape = Signifier)) +
  geom_path(linetype = "longdash", linewidth = 1) +
  geom_point(size=4) +
  scale_x_continuous(limits=c(2012,2023), breaks = 2012:2023, minor_breaks = NULL) +
  scale_color_manual(values = c("darkgoldenrod", "magenta3")) +
  ggtitle("Estimated prevalence of god and jesus signifiers", "within US Twitter users' profile bios
  xlab("Year") + ylab("Prevalence\n(per 10,000 accounts)") +
  theme(text = element_text(size=16)) +
  theme(legend.position = c(0.75, 0.4),
      legend.background = element_rect(fill = "white", color = "black")) +
  labs(caption = "Source: Ipseology - a new science of the self\n \uA9 Jason Jeffrey Jones. You may s
  theme(plot.caption = element_text(size=10, color = "#666666"))
```

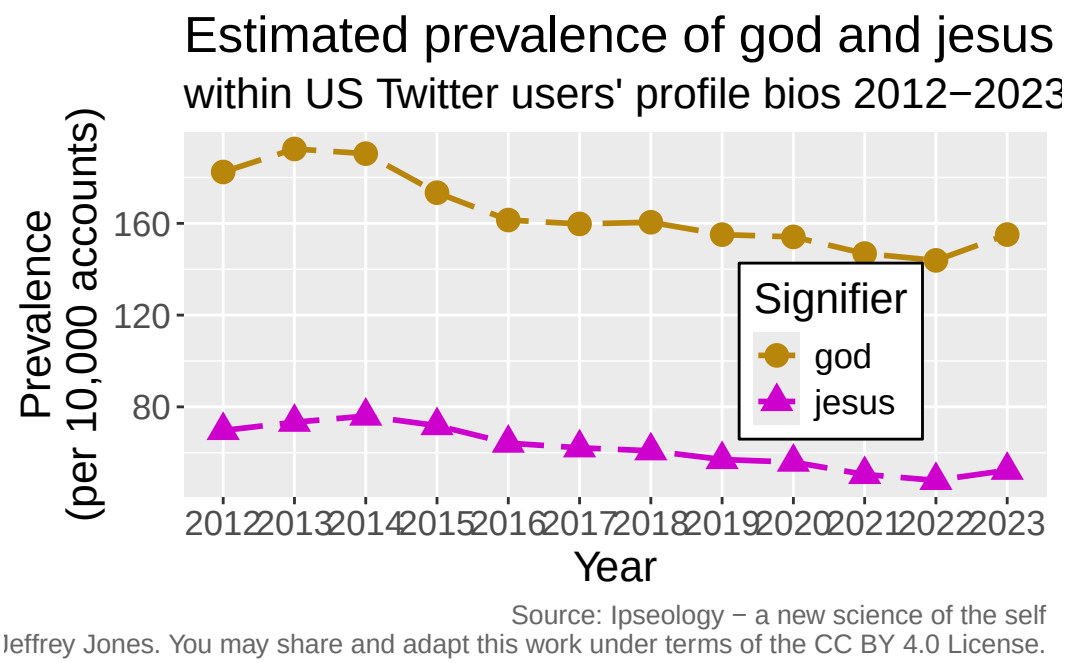


Figure 5.1: Prevalence of god and jesus signifiers in US bios over time.

We see that prevalence of `god` and `jesus` ended at lower values than they began. One would rightfully wonder if users are deleting these words from their bios. However, from cross-sectional analysis it is impossible to know. We must use ***longitudinal analysis*** - in which we observe the same users' bios at different times - to observe and tally bio revision events.

Keeping it simple again, let's consider every possibility for `god` when we observe one user's bio exactly twice.

Table 5.1: Illustrating bio revision events for `god`

| Early Bio | Late Bio | God Events |
|---|---|---|
| All glory to Jesus! | All glory to God! | Add God |
| All glory to God! | All glory to Tom Brady! | Delete God |
| All glory to God! | All glory to God! | Keep God |
| All glory to Belichick! | All glory to Tom Brady! | Ignore God |

There is an important, slightly tricky idea you should internalize before we move forward: Ignore events are extremely common. Most tokens never appear in one individual's bio. There are just too many words in the English language and only 160 characters to present one's self. As a consequence, Add and Delete events are rare (compared to Keep events and especially Ignore events), but they are much more interesting.

## 5.2 Events from Year-over-Year Longitudinal Data

In ipseology, we use year-over-year longitudinal data to observe and tally bio revision events. By *year-over-year longitudinal*, I mean that we match individuals from consecutive years (e.g. 2015 and 2016) so that we can compare their earlier and later bios.

After that, we can do something very interesting: we can count the bio revision events we observe for *every* signifier! We'll stick to `god` and `jesus` for now, but check Data and Code for links to data on all tokens.

In Figure 5.2, compare the Add and Delete counts each year. Now we can make strong inferences about individual behavior in the aggregate.

```
# Download a csv file containing annual events.
# Read about the data in the text file at https://osf.io/57wcb
tyoyl = read_csv("https://osf.io/download/gj8vt/")

godJesusSignifiers = c("god", "jesus")

# Visualize god and jesus Adds and Deletes over time.
tyoyl %>%
  select("token", "lateYear", "adds", "deletes") %>%
  filter(token %in% godJesusSignifiers ) %>%
  mutate(Signifier = factor(token, levels = godJesusSignifiers) ) %>%
ggplot(aes(x = lateYear, y = adds, color = Signifier, shape = Signifier)) +
  geom_line(linetype="solid", linewidth=1.5) +
  geom_point(size=2, stroke = 1.25, fill = "white") +
  geom_line(aes(x = lateYear, y = deletes, color = Signifier), linetype="solid", linewidth=1.5, alpha
  geom_point(aes(x = lateYear, y = deletes, color = Signifier, shape = Signifier), size=2, stroke = 1
  scale_x_continuous(limits = c(2012,2023), breaks = seq(2012,2022,2)) +
  #scale_y_continuous(limits=c(0,6000), breaks = seq(0,6000,1000), minor_breaks = NULL ) +
  scale_color_manual(values = c("god" = "darkgoldenrod", "jesus" = "magenta3")) +
  #scale_shape_manual(values = magaSignifiersShapes) +
  xlab("Later Year") + ylab("Add and Delete\nIncidence") +
  theme(text = element_text(size=14)) +
  theme(legend.position="none") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(caption = "Source: Ipseology - a new science of the self\n \uA9 Jason Jeffrey Jones. You may s
  theme(plot.caption = element_text(size=10, color = "#666666")) +
  facet_wrap(vars(Signifier), scales = "free_y")
```

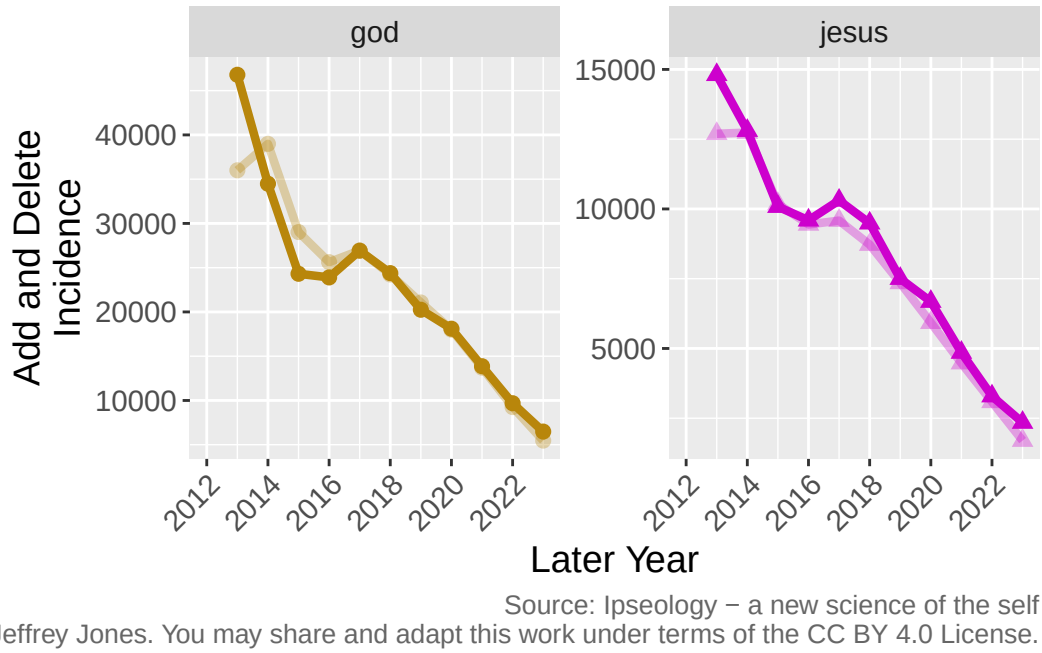


Figure 5.2: Incidence of god and jesus Add and Delete events in US bios within year-over-year longitudinal samples. Stronger lines depict Add event counts, while fainter lines depict Delete event counts. Note that the y-axes have separate scales.

In most years, the number of Add and Delete events were nearly even. There was some separation early in the `god` series. From 2012 to 2013, many more users added `god` (and `jesus`) than deleted. That reversed for `god`, however, which suffered net negative year-over-year periods in 2013-14, 2014-15, and 2015-16.

Overall, the slopes of all four series are negative. This implies that within bio revisions, `god` and `jesus` received less attention each year - with the exception of 2016-17. Think back to the decreasing prevalence of `god` and `jesus` in the cross-sectional data. We can see that it was **not** the result of many more users deleting rather than adding these signifiers to their bios. Instead, there must be other explanations. Perhaps newly joining and active users were less likely to include `god` and `jesus` in their bios. Perhaps those with `god` and `jesus` in their profiles left the platform or stopped tweeting as often. We don't have to guess. Deeper investigation would reveal the paths users took.

I hope you are intrigued. I don't know *why* US Twitter users added and deleted `god` and `jesus` at these rates, but we now know that they did. With the Data and Code section below, you have the resources to quickly compare `god` and `jesus` to about 20,000 other signifiers! I look forward to seeing your comparisons between signifiers of interest to you and to reading competing explanations for the trends depicted here.

## 5.3 Data and code.

Download data for this chapter from https://osf.io/download/cdzsb/ and https://osf.io/download/gj8vt/

R code to produce the numbers and figures within this chapter is embedded in the code chunks above.

# 6 The future of ipseology

## 6.1 The demise and legacy of the Twitter 1% stream

The main source of ipeological data had been Twitter. Specifically, most of my datasets were derived from the stream of a 1% random sample of all tweets. That has disappeared. But all is not lost:

- There is still plenty of science to be done with a decade-plus worth of data from 450 million global users.
- In 2024 and beyond, I am still observing hundreds of thousands of Twitter bios and their edits for users in the US and elsewhere.
- With modest funding, we could collect self-authored self-descriptions through representative-sample surveys.

## 6.2 Further development of ipseological concepts

With students, I am currently conducting further analysis pertaining to bio revision events and the following.

### 6.2.1 What is an *identity alloy*?

In ipseology, an ***identity alloy*** is the mixture of two elements of identity. For example, in the bio *Father of two and prototypical Scorpio*, `father` and `scorpio` form an identity alloy. Every pair of signifiers is an identity alloy. Some alloys are observed more frequently than others.

### 6.2.2 What is an *identity transmutation*?

In ipseology, an ***identity transmutation*** occurs when an an individual stops using one signifier to describe themselves and starts using a different signifier. For example, if I edit my bio from *Scientist who studies identity* to *Ipseologist who studies identity*, then I have transmuted from `scientist` to `ipseologist`.

## 6.3 New directions for ipseology

As computational social scientists and ipseologists, we need identity data **at scale**. I suggest two paths to new data streams. The first is traditional surveys. Scales that elicit personally expressed identity text are sometimes referred to as Who am I? instruments. Instruments like these can be delivered as short surveys. For a low investment of research expenditure ($1000) one could administer an instrument such as the Twenty Statements Test to a representative sample of a few hundred respondents.

Second, I urge the development of web and phone apps to collect self-authored, self-descriptive text from longitudinal panels. An app might simply collect periodic bios from volunteer citizen scientists. A more

ambitious approach would be to build an app that provides value for users – perhaps feedback or accountability on goals for personal growth – in exchange for use of personally expressed identity text in research.

I have become fascinated with self-authored self-descriptions no matter the source. I have become an ipseologist. There is so much opportunity to learn more about our selves; I hope you will too!

-Dr. Jason Jeffrey Jones

# An ipseology glossary

Add (event)

Events are tallied when a bio is observed at two time points. An Add event occurs when the token of interest is not present at the first moment and is present at the second moment.

Alloy

In ipseology, an alloy is the mixture of two elements of identity. If two tokens are present in a bio at the same time, they form an alloy.

Annual resolution

If we observe bios once per year (or sample to one-per-year) we can discuss year-over-year changes/trends in ipseity. Call this annual resolution. Compare to daily resolution.

Bio / Biography

A bio is short text, written by an individual to describe themselves.

Cross-sectional (sample)

A collection of bios sampled over time in which the observed individuals are NOT guaranteed to be the same at each point in time. Cross-sectional samples are useful for describing the active or available population at each moment of observation. Contrast with longitudinal. Reference

Daily resolution

If we observe bios once per day (or sample to one-per-day) we can discuss day-to-day changes/trends in ipseity. Call this daily resolution. Compare to annual resolution.

Delete (event)

Events are tallied when a bio is observed at two time points. A Delete event occurs when the token of interest is present at the first moment and not present at the second moment.

Ignore (event)

Events are tallied when a bio is observed at two time points. An Ignore event occurs when the token of interest is not present at both points in time. Compare to Add, Delete and Keep events.

Incidence

Incidence is a raw count. In ipseology, we might count how many US Twitter users include “mom” or “dad” in their profile biography. It is useful to convert an incidence to a prevalence when comparing across time or location.

Ipseology

Ipseology is the study of human identity using large datasets and computational methods. It is the investigation of ipseity: selfhood, individuality and the elements of identity.

Keep (event)

Events are tallied when a bio is observed at two time points. A Keep event occurs when the token of interest is present at the first moment and is also present at the second moment.

Longitudinal (sample)

A collection of bios sampled over time in which the exact same individuals are guaranteed to be present at each and every point in time. Longitudinal samples are useful for describing the changes within individuals across moments of observation. Contrast with cross-sectional. Reference

Ngram

An ngram is a sequence of tokens. See the definition of Token below. An ngram may be a signifier, and thus, an element of identity. Explore ngram usage within Twitter bios in the United States or multinationally.

Personally expressed identity

Personally expressed identity is who or what an individual themselves says they are. It is personal – the individual is describing themselves. It is expressed – these are words the individual emits, where others might see them. And it describes identity – the explicit purpose of the text is description of the author.

Prevalence

Prevalence is a normalized count. It is the incidence (raw count) of an outcome divided by the potential occurrences. In ipseology, we might wish to contrast how many US Twitter users per 10,000 include "mother" with how many Mexican Twitter users per 10,000 include "madre" in their profile biography.

Signifier

A signifier is a symbol used by an individual to describe themselves. In ipseology, tokens and ngrams used within personally expressed identity text are signifiers.

Social role signifier

A social role is a socially constructed set of expectations regarding behaviors, rights and obligations of a person when they occupy a particular position. A social role signifier is a token or ngram that represents a social role, i.e. a word that denotes a position that most members of a society would recognize and understand.

Token

A token is one linguistic unit. Bios consist of tokens, and we may split a bio into sequences or a set of tokens. Tokens include words such as one would find in the dictionary, but also word-like things such as abbreviations, hashtags and emoji. Crucially, and happily for our purposes, a great many of these tokens describe aspects of identity – such as social roles, affiliations and personal traits. A token may be a signifier, and thus, an element of identity. Explore token usage within Twitter bios in the United States or multinationally.

Transmutation

In ipseology, a transmutation is the change within an individual from one element of identity to another. If Token A is present in a bio at the first moment, and it is not present at the second moment, while Token B was not present at the first moment and present at the second, then the individual has signaled a transmutation from Token A to Token B.

# Annotated bibliography

HINENI: Human Identity across Nations of the Earth Ngram Investigator

The most comprehensive freely available data for the study of human identity. HINENI charts the prevalence of identity signifiers over 32 nations, 12 years, many languages, and millions of individuals.

The evolution of occupational identity in twitter biographies

We compiled the transition matrix between job titles in English language bios over time. Watch as `students` become `engineers` and then `founders`.

Slava Ukraini: Exploring Identity Activism in Support of Ukraine via the Ukraine Flag Emoji on Twitter.

We observed Americans rapidly adding the Ukraine flag emoji to their bios. This was my first work to also examine the *name* profile field. As it turned out, more accounts displayed the emoji in the name than the bio.

Pronoun Lists in Profile Bios Display Increased Prevalence, Systematic Co-Presence with Other Keywords and Network Tie Clustering among US Twitter Users 2015-2022.

Adding pronouns to one's bio is arguably the most pronounced trend in all the US data from 2015-2022. Here we document the trend, report on identity alloys and show that pronouns are clustered rather than diffuse in the follow network.

A dataset for the study of identity at scale: Annual Prevalence of American Twitter Users with specified Token in their Profile Bio 2015–2020.

The ipseological ur-text. My first call to study identity at scale. Introduced public free data and web tools!

Do LGBTQ-related Events Drive Individual Online Disclosure Decisions?

Pre-registered hypotheses regarding LGBTQ Add and Delete events. As-yet-unpublished, because of the wildly ambitious scope.

Using Twitter Bios to Measure Changes in Self-Identity: Are Americans Defining Themselves More Politically Over Time?

Early evidence that US Twitter users were politicizing their identities.

# Acknowledgements & thanks

This material is based upon work supported by the National Science Foundation under grants IIS-1927227 and CCF-2208664.

Vielen dank to the Center for Advanced Internet Studies. The Center provided a fellowship in the Fall of 2023 that allowed me to focus time and attention on this work.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Huge thanks to Jason Jeffrey Jones Productions for believing in this project and supporting it with more dollars than it will ever make back. To every single one of you over there at JJJPro: you're the best! (One of the great things about being my own publisher is I can say anything I want here.)

Finally, *Hi Sophie, Layla, J.J. and Zookie!!! I love you guys!* Thanks for being the best family in the world.