Data Segments & Techniques Lexicon

UPDATED JANUARY, 2016

This Data Segments & Techniques Lexicon was developed by the IAB Data Council.

About the IAB's Data Council:

The IAB Data Council is dedicated to demystifying data usage and control in the interactive advertising marketplace. The Council objective is to enable revenue growth through the establishment of quality, transparency, accountability, and consumer protection in data usage.

A full list of Council member companies can be found here.

In addition to the IAB Data Council, we would also like to acknowledge Joshua Koran, Consultant to IAB; Paul Cimino, Deep Data Solutions; Matt Cohen, OneStep.com; Kurt Ruf, Ruf.com; Matt Schumpert, Cloudera.com, and Nathan Abse, Consultant to IAB.

This document can be found on the IAB website at:
http://www.iab.com/data_lexicon.

IAB Contact Information:
Patrick Dolan
Executive Vice President and Chief Operating Officer, IAB
212-380-4727
patrick@iab.com

Benjamin Dick
Director, Industry Initiatives
Programmatic | Data | Performance
212-949-2432
ben@iab.com

Table of Contents

I. Introduction

Data is the foundation of interactive advertising. And audience segments—subsets of user data signifying specific facts, interests and other attributes—are a key element in improving both advertiser ROI and publisher yields. Audience segments, and the techniques employed to create and use them, are becoming more complex—as real-time algorithmic scoring and mobile devices increase both the quantities and varieties of input data. These and other rising technologies enhance our ability to create and improve the quality of data used to generate audience segments, helping marketers and advertisers to reach a campaign's target audience.

In November 2011, the IAB Data Council published our first Data Segments & Techniques Lexicon, as a useful guide to simplify the buying and selling of interactive advertising. In the years since, there has been yet another revolution in digital technology—most notably the growth of mobile online advertising, often informed by geo-data, and cross-device segmentation. To address these developments, we have updated the original document.

We recognize that the definition of terms about data often differs among the various media companies and data providers offering data solutions. For example, the definition of an "auto intender" will differ by the companies access to input data, type of data inputted and proprietary methods for processing the data used to generate the audience segment. However, we do hope we can arrive at consistency in how we talk about data. This Data Segments & Techniques Lexicon provides an updated standard from the IAB Data Council to specify how the data is gathered, processed and generated into data segments. The definitions below are a tool to help better communicate with media and data partners when enhancing online advertising campaign performance with audience data.

Media planners and marketers should use our revised lexicon as a reference on the language of different kinds of audience data. We suggest you start with the first **graphic**, below—which offers the four basic variables in data collection, processing and generating audience segments, and move on to the detailed examples that follow. Using this common, agreed-upon language will help you improve the creation and use of audience segments in your digital campaigns.

II. Data Generation Systems

The data generation systems can be classified into devices (e.g., Mobile, Desktop, TV, Cars, etc.), what information is sent (e.g., ads, static content, audio, video, etc.) and what information is collected (e.g., ad or page views, survey responses, purchase transactions, etc.). There is also a set of processes and systems that enrich the information collected through data generation systems.

A) Devices

Each type of user device can collect certain types of information (e.g., a mobile phone may broadcast its location at a more definite level than a desktop computer) at different levels of specificity (e.g., a mobile phone is most often used by the owner of that device, while a set-top-box may not know which person is watching a given video, thus aggregates data collection to the household). The user ID associated with each device (or applications running on it) can also vary (e.g., IP address, MAC address, cookie, device ID, login, etc.). Oftentimes the devices are grouped by the screen size.

Below is a list of frequent devices used to generate data used for advertising:

- Wearable (Digital Watch or Personal Fitness Device)
- Car
- Digital Billboard
- Mobile
- Tablet/Laptop
- Desktop
- TV / Set top box

B) Delivery Systems

Below is a list of content delivery systems that deliver content, communication and ads.

- Ad server
- Application and Web servers
- Cell tower
- Content Delivery Network (CDN) data
- Content Management System (CMS) data
- Demand Side Platform (DSP)
- Digital billboard
- Email systems
- Search engines
- Social networks

C) Collection and Storage Systems

Below is a list of systems that collect and store user activity:

- Applications and Browsers
- Billing systems
- Customer Relationship Management (CRM)
- Data Management Platform (DMP)
- Digital beacons
- Operating Systems

- Point-of-Sale systems
- Survey systems
- Website analytics

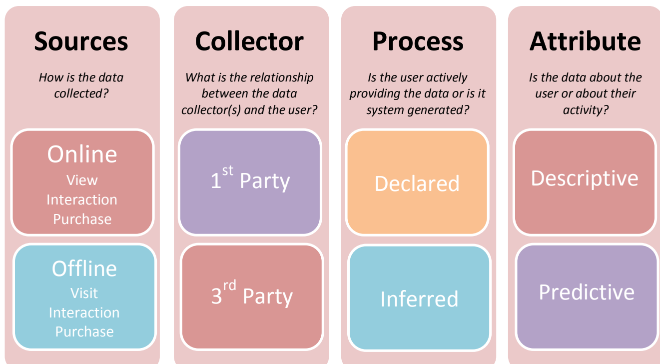
D) Analysis Systems

Below is a list of analysis processes and systems that enrich the information collected by the systems above:

- Audience Segmentation
    - Behavioral Targeting (or interest-based advertising) system
        - Lifestyle
        - Lifestage
        - In store purchase analysis (Recency, frequency, value)
    - Geographic analysis
    - Look-alike modeling
    - Rules-based audience segmentation
- Inventory Analysis
    - Brand safety
    - Fraud detection
    - Viewability
    - Contextual analysis
- Campaign Delivery
    - On target delivery
    - Attribution
        - Cross channel
        - Cross device
        - Online to offline
    - Web site analytics
        - Time spent, session depth, page views, path to conversion
- Creative Engagement Analysis
    - Interactivity
    - Social sharing
- Creative Personalization system

III. Data Segment & Technique Definitions

As seen in our graphic below, there are four broad descriptors—Sources, Collector, Process and Attribute —that define how user attributes are created for use in digital advertising. Every data attribute can be defined by variables in each of the four descriptors. Once data is defined within this taxonomy, it is organized into similar segments, then used to identify and target audiences that match the segment attributes.



A) Sources

How is the data collected?

The source of data refers to both the type of user activity (e.g., view, visit, interaction, purchase) as well as whether this user activity was either digital or in-person. Online activity refers to information generated via digital interactions of the user's device with a remote computer. Offline activity refers to information generated via in-person interactions of the user with a particular physical place. For many years digital activity was solely generated via online interactions (e.g., viewing an ad on a website). However, with the rise of mobile devices, the introduction of digital beacons, and the ability to model / match offline actions to online profiles using browser cookies or device IDs, data is increasingly generated via offline activity. Thus, the source of data could be a combination of both online (i.e. beacon collected data) and offline (i.e. in-store visit) activity.

Below are some examples of the data generated by these systems that fuels the generation of user attributes:

View and Visit

Some types of view information are initiated by the user via a prior interaction (e.g., click on a navigation link), but in relation to ads view information is generally determined by a server. Because of this, a user seeing an ad does not determine what interests a user may have. Whereas the view of a webpage or app often indicates some user interest in the content of that page. It is for this reason that the view information associated with specific products is often used to generate retargeting attributes. View information often specifies the time of the view and the asset the user interacts with (e.g., ad ID, URL, app ID). Depending on the device on which the content is view, the location of the user may also be collected.

Online Examples:

User views the sports page on Yahoo!.
User views a banner advertising running shoes displayed on Yahoo!.
User views a jogging app on their mobile device.

Offline Example:
User visits Macy's.
User visits the shoe department at Macy's.

Interaction

Unlike some types of view information, users must always actively generate interaction data. This conscious choice on the part of the user, makes this type of data more valuable in understanding user interests. Interaction information often specifies the type of interaction (e.g., click, add to cart, tweet, like), time of the transaction and the asset the user interacts with (e.g., ad ID, beacon, social API). In the case of a form (e.g., registration or survey), the information can contain an array of user-declared information.

Online Examples:
User clicked on a banner advertising running shoes displayed on Yahoo!.
User searched for running shoes at Zappos.com.
User tweeted to a friend a picture of the pair of running shoes at Zappos.com.
User added-to-cart a pair of running shoes at Zappos.com, but did not yet purchase.
User filled out a customer survey on Zappos.com.
Offline Example:
User filled out a customer survey at The Sports Locker.

Purchase

Purchase is a particular type of user interaction, which is classified separately due to its great value to retailers. The purchase information often specifies the name of the time of the transaction, the store from which the item was purchased, the cost of the item, the item purchased, and the category of the item. Depending on how the data is collected, the contents of the shopping basket (i.e. the actual items purchased) may be aggregated together (e.g., when the purchase information is collected by a credit card).

Online Example:
User purchased running shoes from Zappos.com.
Offline Example:
User purchased running shoes from The Sports Locker.
User redeems coupon at their grocery store.

Historical Interactions and Purchases

Some data, most frequently from offline sources, contain information that was collected from a user's prior activity (e.g., home purchase, credit application, auto purchase, voting records, government records). Due to regulations oftentimes this information may not be associated directly with the user, but rather associated to a broad set of users based on where they live.

Offline Example:
User lives in a zip code with high percentage of home ownership, according to public tax records.

B) Collector

What is the relationship between the data collector(s) and the user?

The collector of data refers to the entity that gathers and stores the user activity and derived information associated with that information. Oftentimes the attribute eventually generated from the collected information is derived from multiple different collectors.

One of the key principles in defining ownership and control of data is determined by the relationship between the data collector and its user. There are three forms of relationship: 1st party, 2nd party and 3rd party.

1st Party

A "first party" is an entity that collects information from or about users and is the owner or controller of the website or service with which the user interacts directly. A first party also includes the first party's "affiliates."[1]

Examples: The publisher of a site visited by a user —or an advertiser's site the user clicks through to. Data collected and used by the site is first-party data.

2nd Party

A "second party" is a first party that that sells or shares data to a non-affiliated website or service. Given that the rules around data ownership, use and control are governed only in relation to first and third-party definitions, the reason to distinguish a second party from either a first or third party has fallen out of favor, since in relation to data collection it is treated as a first-party and in relation to data sharing it is treated as a third-party. For this reason we did not mention it in the 2011 version of this Lexicon. However, for completeness sake, we wanted to ensure consistent usage of this term.

Online Example:

> *AOL/Adap.tv sells behavioral segments collected from its own website to monetize traffic on Yahoo!.*

Offline Example:

> *Safeway offers discounts on fuel to customers of Chevron to incent users to opt-into rewards program.*

3rd Party

A "third party" is an entity that collects information from or about users from a non-affiliate's website or service. Third-parties, such as data aggregators and ad networks, often create data products that span collection from websites and stores not owned or controlled by a single entity. By aggregating this information, third-parties can offer smaller websites and stores that do not have the technical, data or service resources the ability to compete against large vertically-integrated companies.

---

[1] An "affiliate" is an entity that controls, is controlled by, or is under common control with another entity. Control of an entity means that one entity (1) has significant common ownership or operational control over the other, or (2) can exercise a controlling influence over the management or policies of the other entity. In addition, for an entity to be under the control of another—and thus be treated as first party under these Principles—that entity must adhere to Online Behavioral Advertising policies that are not materially inconsistent with the other entity's policies.

Online Example:

Google Analytics collects a user's visit path when visiting a sports website.
Oracle/Bluekai collects information from ESPN.com.

Offline Example:

Experian collects information about a user's mortgage from their lending institution.

C) Process

Is the user actively providing the data or is it system generated?

The process of generating an attribute can be defined whether the user actively supplies the information (i.e. "declared" data) or whether it is generated from a system (i.e. "inferred" data). The system can either manually create the attribute (e.g., with Boolean logic) or algorithmically generate the attribute (e.g., look-alike modeling). These combination of attributes (e.g., gender and age attributes) forms a "segment."

Declared Data

When a user actively volunteers information such as on a registration form this is called "declared" data. Often this type of data relates to demographics (e.g., age, gender, presence of children, marital status, preferred language, but can also include interests). Declared data is often assumed to be of higher quality, since the user is stating what information they want the data collector to associate with them. However, some users are not truthful in filling out forms (e.g., lying about their income on a dating site), so that merely knowing that the information source is declared is not alone enough to know it is of high quality. To measure the accuracy of both declared and modeled data, a subset can be verified via offline matching and social verification (e.g., few people who lie about what year they graduated a particular school on classmates.com, since the purpose of the site is to connect schoolmates). Given the extra cost involved, often marketers' rely on the campaign results which targeted that particular segment to measure whether this audience data gives a lift in the advertiser's success metrics.

Inferred Data

When a system assigns information to the user based on their activity this is called "inferred" data. For example, if a user visits a baby supplies website the system may assume the user has a baby. However, this inference may be incorrect in the case that the user is actually a grandparent or friend purchasing supplies for the true new parents. Accordingly, systems usually look for repeated patterns of activity to increase the accuracy of their inferences.

A key ingredient in generating accurate inferences from user activity is content classification. Whether the user visits a website or conducts a search, understanding the meaning of that interaction requires disambiguating the meaning of the content to infer the user's intent. For example, if the user enters a search phrase "China pottery" the user could refer to pottery from the country or to porcelain dishware. By processing multiple events from the same user it may be possible to improve the understanding of the user's interest (e.g., a subsequent page or search included information about "importing" or "directions to a local retailer").

Note, the outputted attribute can be the identical type of information as the inputted data. For example, a user could declare they are a twenty-year-old male or a system could infer their gender is male and their age (e.g., given a number of activities normally conducted by college students).

Age of Data

The recency of activity is also important is assessing the usefulness of data. This matters for both declared and inferred data. For example, a user who declared they liked a given musician years ago may no longer represent a current preference for that musician. As another example, a user who visited multiple baby websites many years ago, is likely now to have older child and no longer be interested in baby items. The process of "refreshing" data – updating a segment to ensure that its user pool still aligns with the original descriptive or predictive attributes – is one way that data accuracy is maintained.

Manual vs Algorithmic Generation

User attributes can be assigned both through automated associations via an algorithm process or manually created via Boolean-logic (if-then-else) rules. One standard approach to segmenting users is retargeting, which defines particular user actions with Boolean logic (such as visiting a site but not purchasing within a specific time frame) to assign segment membership. Another standard approach to segmenting customer is look-alike modeling, which finds users based on similar activities and characteristics to another set of users who a marketer has defined as a good set).

D) Attribute

Is the information about the user or their activity?

The data attribute refers to facts describing the user or what they do. Some facts are slowly changing (such as marital status, presence of children, postal address, etc.), while other facts change rapidly (age, income, etc.). Some facts describe what the user does or observes (e.g., visits a page, sees an ad, clicks on an ad, purchases a product).

Descriptive vs Predictive Attributes

Multiple facts about one user are often combined to assign membership into an audience segments (e.g., in-market auto intender, frequent clicker). Thus, if a user visited a series of auto sites this could be used to assign membership into either an auto enthusiast segment (describing past or current state) or an auto-intender segment (predicting a future action). To predict the future action, some process must be applied to the raw facts to infer the subsequent action. Thus when the set of facts are used by themselves we call this a "descriptive" segment and when an algorithm associates these same facts with some other user action we call this a "predictive" segment.

Most often descriptive segments are based on information about the user (e.g., age, income, location, marital status, presence of children). Information about the user's interests is often predictive, since this infers the likelihood of the user's interest continuing into the future, rather than merely describing the user's past interest.

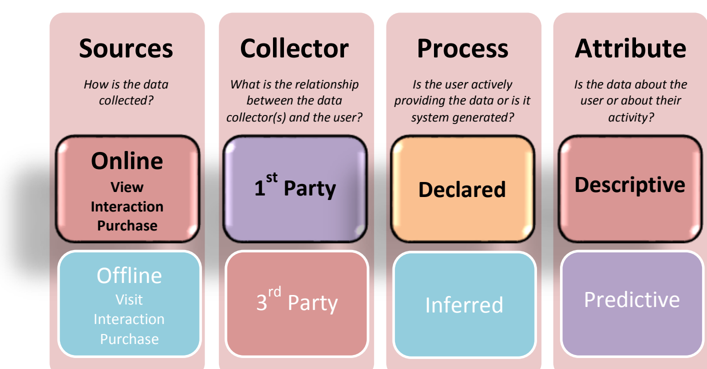
IV. Additional Examples

The following examples illustrate how media and data partners can use the IAB Data Segments & Techniques Lexicon as a communications tool. While not comprehensive, the list provides a sampling of data segments and techniques currently used in online advertising.

Example	Source	Collector	Process	Attribute	Process	Attribute
A	**Descriptive Segment**	**Demographic Attributes** • Gender=Male • Age range= 18-25	*Online*	*1^{st} Party*	*Declared*	*Descriptive*
B			*Offline*	*1^{st} Party*	*Inferred*	*Descriptive*
C			*Offline*	*3^{rd} Party*	*Inferred*	*Descriptive*
D			*Offline*	*3^{rd} Party*	*Declared*	*Descriptive*
E		**Psychographic Attributes** • Political Affiliation	*Online*	*1^{st} Party*	*Inferred*	*Descriptive*
F			*Offline*	*3^{rd} Party*	*Inferred*	*Descriptive*
G	**Predictive Segment**	**Interest Attributes** • Business Traveler	*Online*	*1^{st} Party*	*Inferred*	*Predictive*
H			*Online*	*3^{rd} Party*	*Inferred*	*Predictive*
I			*Offline*	*3^{rd} Party*	*Inferred*	*Predictive*

A) Descriptive Segment (Declared)

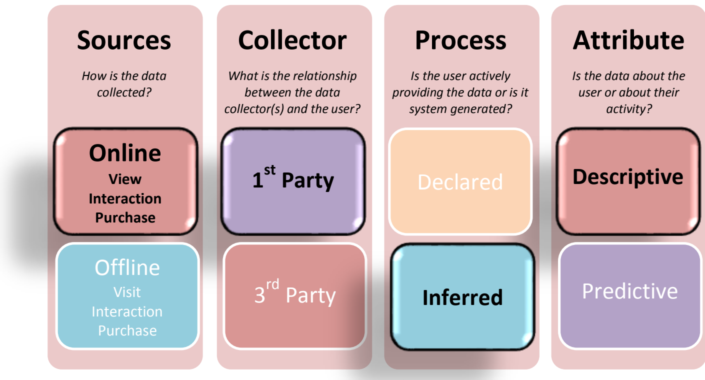
Demographics (Gender=Male, Age range= 18-25)



In this example, a social media company has collected registration information from its users, so the source is online, the relationship of the collector to the user is 1^{st} party and the process of producing the attribute was declared. The actual segment (a combination of gender and age attributes) is descriptive of the user, since it is about them rather than what they are likely to do.

B) Descriptive Segment (1st Party & Inferred)

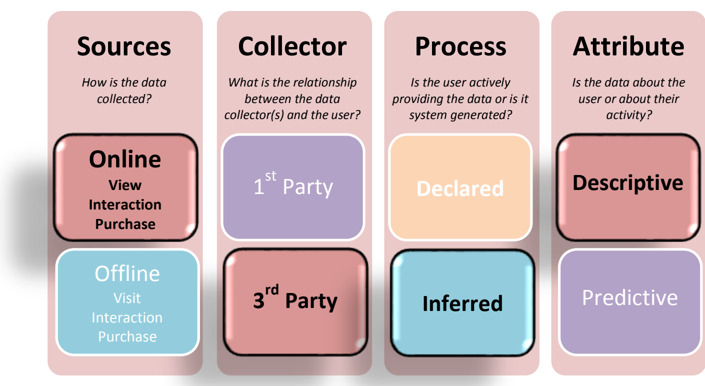
Demographics (Gender=Male, Age range= 18-25)



In this example, a publisher has used modeling techniques to generate a segment of users based on their multiple visits to multiple websites that frequently attract young, male visitors. The source is online, the relationship of the user to the collector is first-party and the process of assigning the attribute is inferred. The outputted segment is still descriptive as it relates to demographics.

C) Descriptive Segment (3rd Party & Inferred)

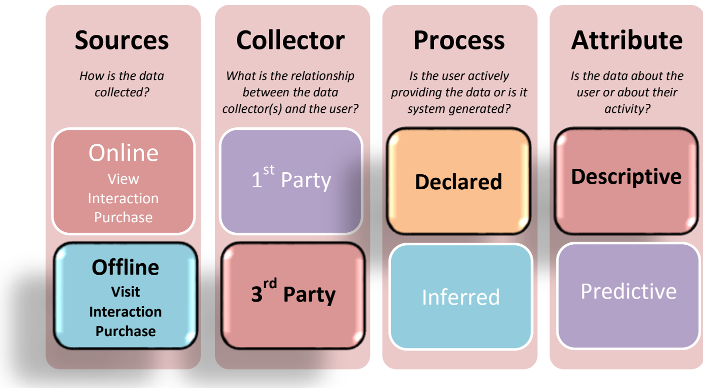
Demographics (Gender=Male, Age range= 18-25)



In this example, a data aggregator has used modeling techniques to generate a segment of users based on their multiple visits to multiple websites that frequently attract young, male visitors. The source is online, the relationship of the user to the collector is third-party and the process of assigning the attribute is inferred. The outputted segment is still descriptive as it relates to demographics.

D) Descriptive Segment (Offline to Online)

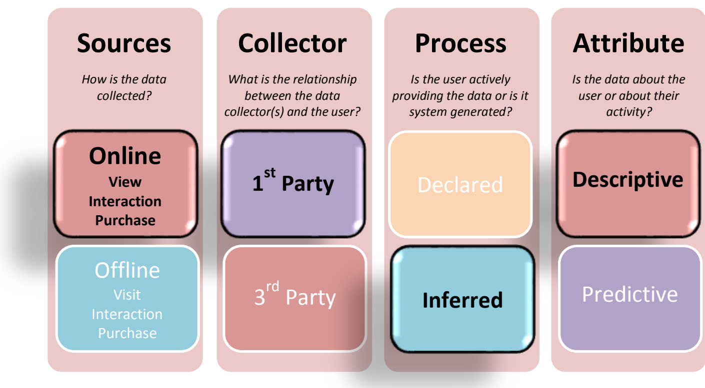
Demographics (Gender=Male, Age range= 18-25)



In this example, a data onboarding company sells to an advertiser the segment based on having matched online user IDs to offline retailer loyalty rewards database (i.e. CRM data). The source was offline, the collector has a 3rd-party relationship to the user, the process is still declared since the user registered to enter the loyalty rewards program and the output segment is still descriptive.

E) Descriptive Segment (1st Party & Inferred)

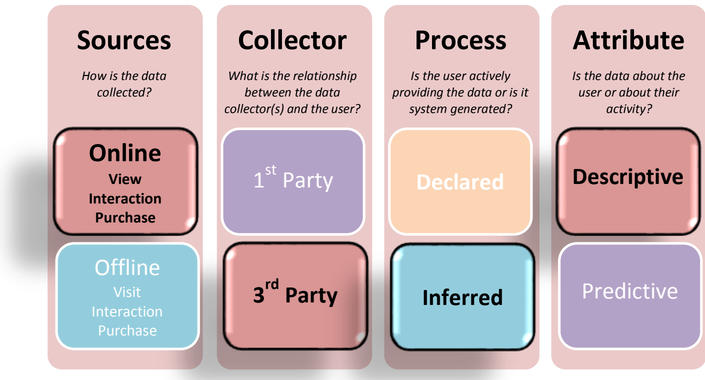
Psychographics (Political Affiliation=Conservative)



In this example, a news publisher has used modeling techniques to generate a segment of users based on their multiple visits and which articles they've read and search terms they've entered to infer that the users have a conservative political affiliation. The source is online, the relationship of the user to the collector is first-party and the process of assigning the attribute is inferred. The outputted segment is still descriptive as it relates to psychographics, rather than future behavior.

F) Descriptive Segment (3rd Party & Inferred)

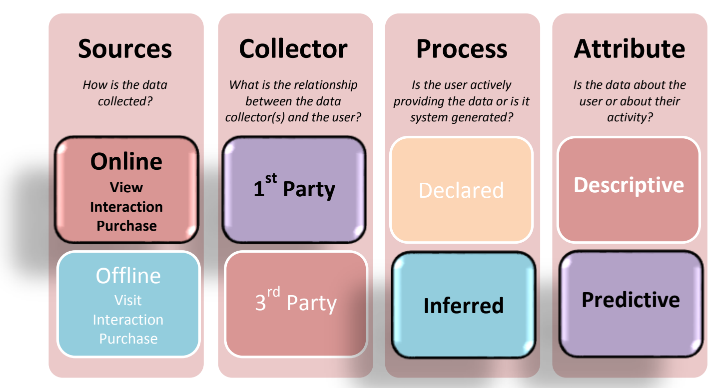
Psychographics (Political Affiliation=Conservative)



In this example, an ad network has used modeling techniques to generate a segment of users based on their multiple visits to a single site and which articles they've read and search terms they've entered to infer that the users have a conservative political affiliation. The source is online, the relationship of the user to the collector is third-party and the process of assigning the attribute is inferred. The outputted segment is still descriptive as it relates to psychographics, rather than future behavior.

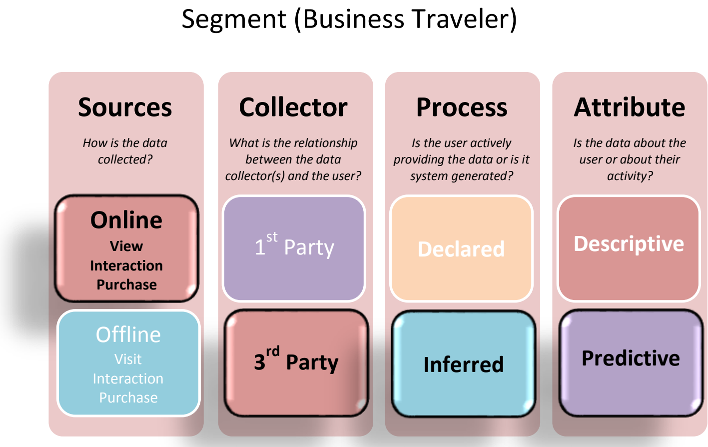
G) Predictive Segment (1st Party & Inferred)

Segment (Business Traveler)



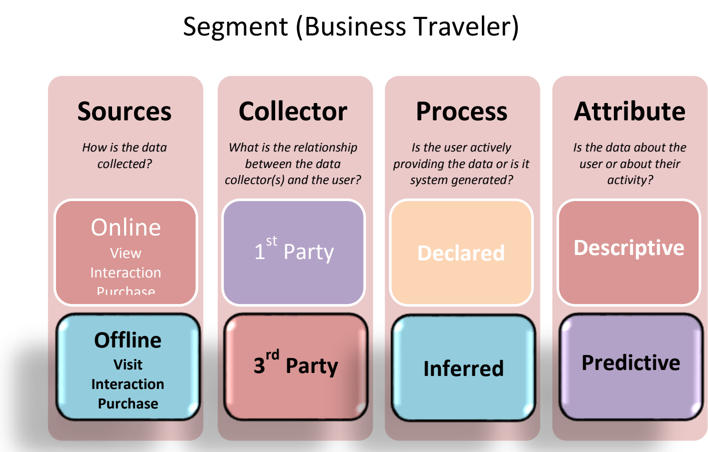
In this example, a travel comparison publisher has used modeling techniques to generate a segment of users based on their frequent reservations on airlines to infer that these users have been and will continue to travel for business. The source is online, the relationship of the user to the collector is first-party and the process of assigning the attribute is inferred. The outputted segment is predictive, since the attribute describes the future behavior of the users in the segment.

H) Predictive Segment (3rd Party & Inferred)



In this example, a data aggregator has used modeling techniques to generate a segment of users based on their frequent reservations on airlines collected across multiple travel comparison websites to infer that these users have been and will continue to travel for business. The source is online, the relationship of the user to the collector is third-party and the process of assigning the attribute is inferred. The outputted segment is predictive, since the attribute describes the future behavior of the users in the segment.

I) Predictive Segment (Offline & Inferred)



In this example, a data aggregator relying on a network of beacons has collected information from multiple car rental agencies and airports and assigned this information to anonymized mobile IDs. The source is offline, the relationship of the user to the collector is third-party and the process of assigning the attribute is inferred. The outputted segment is predictive, since the attribute describes the future behavior of the users in the segment.