Methodology of gemiusAudience study

for Gemius clients and salespersons

October 23rd, 2023

Measuring digital and beyond

Gemius since 1999

Table of content

Preface 3
1. Introduction 3
2. Intuition and short definitions.....3
3. Overview and scheme of gemiusAudience study.....13

Preface

This document is intended as an easy and intuitive, yet complete, description of the gemiusAudience study and its position among other Gemius products. It might be given to the clients, used as an appendix to contracts, and also as training material for salespersons. It does not contain too specific information, which might be potentially used to recreate Gemius's methodology by competitors, so it is safe to share it.

1. Introduction

gemiusAudience is the longest existing Gemius product, which was the center of Gemius offer for many years. We might say that it delivered the foundation of knowledge and experience, required to build the next generation of products like gemiusAdReal, gemiusDirectEffect, and gemiusPostBuy. Because of that, these products share many definitions, data sources, measurement scripts, and methodology with gemiusAudience. So, this document, in fact, helps understand not only gemiusAudience per se but the majority of Gemius's methodology.

2. Intuition and short definitions

Site-centric data (aka "universes") – data coming from direct measurement on the publisher website, stream, or application, done by Gemius measurement scripts. They consist of explicitly measured metrics like Views (page views, stream views), Visits, and Time, but also of Real Users number, which is an estimate calculated later using advanced methodology. We treat this site-centric measurement as a reference, to which we later adjust panel results by weighting or performing other bias-removing algorithms. Products that currently use these site-centric data as reference are: gemiusAudience, gemiusAdReal, gemiusDirectEffect, and gemiusPostBuy.

User-centric data – common name for all panels (i.e., panelists' activity and demography).

To be more precise: data comes directly from the users' side (cookie/software/hardware panelists). Panelists' demography (answers in Gemius' questionnaire) is combined with their activity on websites, players, and applications.

The activity of cookie panelists comes from the site-centric, activity of software panelists - from Gemius' browser extension (for PC) or Gemius' mobile application (for phones), and activity of

hardware panelists - from Gemius' measurement-optimized smartphones. More information on panel characteristics is provided below.

Sonars – a method of measuring time spent by internet users on browser's active tabs, which is embedded into Gemius's scripts and which main goal is to perform accurate measurement with possibly the lowest number of measurement hits. Scripts decide every second if they should send a "alive signal" for a given audited webpage open in a tab or not, and the decision is made with low probability (usually 1/40 or 1/3000 depending on the market). In this approach, the number of needed hits is low, and measurement for a single cookie is not accurate enough (especially at 1/3000 probability), but for all cookies visiting major websites, the method has very good statistics and is accurate.

Auto-refresh filtering – a process of removing hits incompatible with the Gemius page view definition (an intentional user action) that occurs as a result of periodic site content refreshing (e.g., updates of sport results during the game). Filtering is based on analyzing the time intervals between consecutive hits from the same user on the same URL. When a given exact time interval happens unnaturally often (at least 2x more often than the median occurrence from 60-time intervals around), it is marked as an auto-refresh and corresponding hits are filtered out.

Cookie Panel – a group of cookies, to which we displayed the Gemius questionnaire (while a cookie was visiting an audited website) and the users answered all the questions. For every cookie in Cookie Panel, we know its declarative demography and we record all its activity on audited websites and players. We do not have any information about non-audited websites coming from Cookie Panel.

Software Panel – a group of panelists who, after filling in the questionnaire, also installed specialized Gemius measurement software on their PC, Tablet or Phone. This method of measurement has wider possibilities than Cookie Panel because it can measure also non-audited websites, players, and applications, but it has a much lower conversion rate and because of that Software Panels are much harder to gather and smaller, thus giving higher statistical error.

Single Source Panel (also known as Hardware Panel or Device Panel) – a research panel, in which Gemius signs long-term agreements with panelists and gives them measurement-optimized smartphones, capable of measuring their TV and radio activity, thanks to so-called "audio-matching" (see definition below).

Panelists are also obligated to install measurement extensions on their PC browsers (i.e., become Software Panelists), so we can also track their PC traffic. Consequently, we analyze reach on the internet (smartphone and PC), radio, and TV, all together in one panel.

Gemius's strategy is formulated as one sentence: the single source for digital media. Gemius' Single Source Panel is the foundation and the direct embodiment of this strategy.

Audio-fingerprint – a numerical code generated in a lossy way from audio recording, which is precise enough to allow comparison if two different recordings have the same content but makes it impossible to recreate or interpret the original sound.

Reference Signal – a set of audio streams from live TV and radio channels, which are the reference to which we compare and fit audio-fingerprints from measurement-optimized smartphones in the Single Source Panel.

Audio-matching – a process of finding look-alikes between audio-fingerprints from measurement-optimized smartphones in the Single Source Panel and the Reference Signal. The aim is to detect which TV or radio channels the given panelist has heard. Audio-matching is made in a very safe way, with full respect for panelist privacy, so it is impossible for Gemius to obtain any other information from audio-fingerprints, than a fact of hearing radio or TV programs and ads.

Gender Automatic Prediction Algorithm (GAPA) – an algorithm that aims to predict panelist gender, based on the share of websites and applications, which are used predominantly by women or by men, and which are present in panelist's activity. Reference data about gender shares, come from Phone Cookie Panel and Phone Software Panel. GAPA is used for Single Source Panel quality control: in the case of a mismatch between declared and predicted gender, a special procedure starts, to verify if the panelist abides by the terms of the contract with Gemius, mainly if she or he did not pass measuring smartphone to someone else. Using the same principles, GAPA might be used to find activity typical to children and find a mismatch, if an adult was declared as the primary user of a given phone. The algorithm uses a modest set of data about panelists (only visited domains and apps) and is very simple in its construction (e.g., no machine learning is used). That is why it is designed to find only pronounced cases of mismatch (i.e., outliers) and it cannot be used for universal gender prediction.

Required panel size – one of the most basic rules of statistics is: "the larger the sample, the better the results". Classical statistics let us calculate the sample size (i.e., number of panelists) needed to achieve the given limit of error. To do that, we need to define the margin of error, confidence level (i.e., probability that actual error will fit in the margin of error), and population size.

Let's take an example: population of 2 000 000 people, margin of error ±3pp (percentage points or percent of population), confidence level of 95%. For such input parameters the required sample size is 1067 panelists (there are dozens of easy-to-use online calculators for that, so we do not provide

equations). Important note: this number does not increase with increasing population, so for any larger population it is still 1067!

In such a classical approach we assume that our sample is completely random and thanks to that very representative. But if our panel is not ideally representative and we need to use a weighting procedure, we have to divide the required panel size by weighting efficiency (definition below). The lower the weighting efficiency, the bigger the size of a panel we need in order to correctly represent all target groups. Let's say that we weighted the panel, obtaining weighting efficiency of 40%. This means that we need to perform the calculation: 1067/0.4 = 2667.5. So, the end result for our required panel size is 2668 panelists, for us to be 95% sure, that our error fits into ±3pp range.

Population data (aka Structural data) – an establishment survey, a dataset describing the demographic structure of a given population (i.e., percentage of men and women, age distribution, place of living, etc.).

Good Cookie (in general, **Good Identifier**) – a cookie active at least once: before, during, and after analyzed period. If we see good cookies before and after the analyzed period, we are sure that they were alive during the whole analyzed period. The basic assumption is that these non-deleted cookies (i.e., good cookies) are statistically representative of the whole market.

BEAST (Browser Estimation Algorithm Standard) – a predictive algorithm, giving us the probability that a given cookie will turn out to be a good cookie in the future. Prediction is made based on the level of previous cookie' activity and the time that passed since last day we saw this cookie.

OverNight – a part of the current production process of gemiusAudience study that uses

- BEAST methodology in order to calculate results the next day (just over the night) for the previous day/week/month, and
- BPS process in order to find deduplication between platforms,

and then present deduplicated RU for combined platforms like PC (PC-home & PC-work) or Total (PC & Mobile).

AppVisitorID – Phone, Tablet, and CTV apps might deliver various identifiers, with different levels of restrictions and we simply want to always use the best identifier available. We rank them by usefulness and quality they give to our study:

1. Advertising ID – the best, common for all applications on a given device, but will be probably eliminated in the near future by OS developers,

2. Vendor ID – hard to change by a user, but does not provide duplication between apps of different vendors,
3. Standard application cookie – also does not give duplication and is often the easiest to delete (resulting in considerable rotation).

From a technical point of view, to make it easier to work with, we decided to create a new identifier, which consists of the best currently available identifier from the three ones mentioned above and we call it “AppVisitorID”.

EC (Estimated Cookies) – an estimated number of cookies that would be registered on a given node, if there is no cookie deletion at all and all browsers accept Gemius cookies. In such a hypothetical situation each browser would be represented by one cookie. In other words, EC is an estimated number of browsers that visited a given node in a given period of time.

EC Standard – a formula for the estimated number of browsers where the scaling factor is based on considered node n. **The method is recommended since 2023.**

$$EC_{Standard} = GC_n * \frac{AH_n}{GH_n}$$

EC Global – a formula for an estimated number of browsers where the scaling factor is based not on considered nodes, but the whole audited internet. Widely used so far, but should be changed to EC Standard, which is our current recommendation.

$$EC_{Global} = GC_n * \frac{AH_{Internet}}{GH_{Internet}}$$

J_BRUS – estimated number of users per browser, calculated per node, per platform. It depends on two factors: market-specific saturation of devices (so-called J structural) and relative size of the analyzed node. Bigger nodes have J closer to the number for the whole internet and smaller nodes have J closer to 1.

Real Users or **RU** – estimated number of real people (not computers, cookies, or IP addresses) who visited a given node. It is calculated as the number of browsers (EC) times the average number of people per browser (J).

$$RU = EC * J$$

Real Users for Mobile Apps – analogically to the definition above, it is an estimated number of real people using Mobile applications. The method also uses the same equation as above.

However short, additional comments are needed:

1. The recommended identifier to use while calculating apps' EC is AppVisitorID,
2. We assume that all AppVisitorID are "Good Identifiers", so we do not need to use the BEAST method,
3. We use the J_BRUS method, where J and EC for the whole Internet come from Phones WWW data, as the best available approximation.

Reach of the study – a percentage of internet users on a given market, who visit at least one website audited by Gemius per month.

PRES (Population and Reach Estimated Smoothly) – a method of estimating the population for every platform and the reach of the study on a given market.

The basic rule of PRES methodology for population estimation is to use an external study as a starting point and then introduce into it a trend analogical to monthly EC changes.

The reach of the study is estimated directly from the Software Panel, if it is available in a given market. If only the Cookie Panel is available, then we teach a machine learning regression algorithm on Software Panel data (from all the countries where it is available), to predict the reach based on characteristics of the Cookie Panel.

Home/work correction – many people use different PCs at home and at work even during the same day. That is why we need to know the duplication between these two groups of computers, in order to get correct (deduplicated) Real Users numbers for the whole PC platform.

H/W classifier is a machine learning classifier, that recognizes whether a given IP corresponds to a home or work address and then divides traffic on the PC platform accordingly.

In our PC questionnaires, we ask panelists whether they are at home or at work while answering our questions. Thanks to that, we gather a learning set of home and work IP addresses. Then, we teach a machine learning classifier to use this set, in order to assess all other IP addresses observed in a given market.

Facebook/non-Facebook correction – Facebook phone and tablet application might serve as an internet browser (called WebView or in-app browser), but it does not share cookies with any regular browser. Because of that, we might observe on some nodes two cookies coming from one person: one from the regular browser, and one from the Facebook application. This might lead to an overestimation of RU on the nodes popular on both of them.

Facebook/non-Facebook correction is a method of splitting cookies between two technical platforms called Facebook and non-Facebook, and then estimating and introducing duplication between them. Thanks to that, we predict what average percentage of cookies from both platforms (i.e., Facebook and non-Facebook) were in fact generated by the same people, we take this into account and correct the final results accordingly.

It is a custom-made method for this one specific application. But in the future, there might be some other applications with significant in-app browser traffic and their own cookie space. Currently, we do not have a general solution for all of them, but it is planned.

RIM weighting (Random Iterative Method) – a process of assigning a number called “weight” to every panelist, to determine how many people from the real population they should represent, and as a consequence to level off all the biases that might occur in our panel. For example, if we know from a structural study, that there should be 50% of women and 50% of men in our panel, but instead we see 40% of women and 60% of men, then we simply assign higher weights to women panelists, so together they sum up to 50% of all panelists’ weights. When we remove biases concerning structural conditions, like gender, age, education, etc. it is called “structural weighting”. We also might remove biases regarding panelists' activity, e.g., if we have too low number of panelists active on a given node, we need to increase the weights of these panelists to meet reference conditions (RU from universes) and then it is called “behavioral weighting”. Weighting is a very common process in our data flow. In the standard gemiusAudience production it is used dozens of times to make sure that on many different stages of production, the panel dataset always meets certain conditions.

Weighting Efficiency (WE) – a measure expressed in percent, which is meant to describe how uniform the weights of panelists are. It is a very desired property of the weighting process, to give weights as uniform as possible, because it assures smoother and more accurate results on small nodes. Weighting efficiency equal to 100% would mean, that all the weights are exactly the same, and weighting efficiency close to 0, that weights come from an extremely wide range.

In Constant Panel (detailed description below) all panelists have equal weights, so we might say that technically it has WE=100%. But Constant Panel is a virtual, modeled panel, which is not directly weighted, so, in this case, it does not make any sense to talk about weighting efficiency. We simply do not define it for Constant Panel.

Census Safeguarding Age Correction (aka "Central Statistical Office correction") – a set of weighting conditions regarding age (or age-gender crossings), which go beyond standard structural data and are based on the official census for a given country. They are constructed in such a way that the sum of panelists’ weights at a specific age and gender cannot exceed (by e.g., more than 3000) the number of people corresponding to a specific age and gender in the whole population of the country.

These are solely safeguarding conditions, so if the weighting process goes normally, they do not change anything. They come into action only in extreme and rare cases, when standard weighting,

because of strong panel bias or extremely low weighting efficiency, would return illogical results of some narrow age group, exceeding the number of people at this age in the population. Then, additional conditions reduce weights for this narrow group and re-distribute them to other ages within the same structural age range, as a result keeping both the standard, structural conditions and not exceeding census.

AB merging (Application Browser merging) – a method of assigning mobile application activity to mobile Cookie Panel, which normally does not have this activity (we do not recruit panelists on mobile applications.) The main idea behind the method is to join application identifiers (so-called AdvertisingIDs) with cookie panelists from a standard browser, by checking if these two identifiers appeared under the same IP at the same time.

Balance Amendment – a method complementary to the structural and behavioral weighting of a panel. Designed to remove a potential systematic estimation error of non-audited nodes, caused by panel bias (i.e., too high or too low average panelists' activity). The method automatically finds the optimal RU correction function that gives us the lowest average RU error first on audited websites (where we have both site-centric and weighted panel data to compare to). Then this correction function is applied to non-audited nodes to minimize errors on them. The method also has a customizable safeguarding mechanism built-in which further reduces the probability of overestimation on non-audited nodes.

Cookie-Software Fusion – a process in which we combine information from two panels. Cookie Panel gives us information about audited websites and Software Panel about non-audited ones. The Cookie Panel is much bigger and thanks to this of better quality than the Software Panel. That is why we treat the Cookie Panel with priority and attach to it information about activity on non-audited websites from the Software Panel. Optimal connections are found as look-alikes, i.e., we connect a Software Panelist to a Cookie Panelist with the most similar internet activity. It is worth noticing that activity on audited websites from Software Panel, after finding optimal look-alike, is discarded (because Cookie Panel already delivers a better version of such activity).

BPS (Behavioral Panel Synthesis) – a method for merging panelists from different platforms (PC, Phone, and Tablet), into one panelist, who now might have cross-platform activity. The algorithm of panelists merging is mainly based on behavioral look-alikes, i.e., we want to connect panelists who visited a given node on PC with panelists who visited this node on Phone. As a consequence of this approach, BPS introduces duplication of activity between platforms. Its final aim is to recreate duplication observed in the so-called calibration panel (group of panelists active on more than one platform in the same time period). The main purpose of the BPS method is to calculate the combined audience of a given node, i.e., Total Real Users, no matter which platform they came from.

Constant Panel – a modeled panel which contains a very high number of virtual panelists with equal weights. Instead of changing the weight of one panelist, we simulate - a larger number of panelists from large target groups and a smaller number of panelists from small target groups. The panel is called “constant” because the number of panelists and their weights do not change in time.

By using a non-rotating set of panelists with equal weights, we directly simulate the situation of a real-world population, where all people are counted equally, and natural rotation is quite low. Thanks to this stability, we can easily calculate consistent results for any selected custom time period directly from the panel.

Constant Panel might be formed from any real input panel, and also from many real panels, if each one of them provides a different part of the panelist activity. On the markets, where it is available, Constant Panel consists of the Single Source Panel as one of the inputs, providing unprecedented quality of cross-media measurement, combining internet, TV, and radio data.

Daily fusion – a process of mapping panelists from results of yesterday into the previous 27-day constant panel, reaching the target of 28-day RU value coming from the BPS process, for each node and platform simultaneously.

FPA (Fill Panelist Activity) – an algorithm that edits panelists' activity to make it match exactly with the measured site-centric results (Views, Visits, and Time), by adding or removing activity on given nodes. One might say that it could be done by weighting process instead, because the goal is similar, i.e., to match site-centric reference results. But the weighting assigns different weights to different panelists, changing the general shape of all metrics, including the Real Users metric, and making them only roughly equal to the ones from the site-centric. FPA instead is a fine-tuning algorithm that precisely modifies small parts of panelists' activity, already after weighting, to get much more precise results mainly on Views, Time, and Visits, while trying not to change Real Users metric.

Heavy Panelists Filtration (HPF) – a classical outlier removal method which aims to find and minimize abnormally high values of Views, Visits, and Time coming from the activity of a single real panelist. HPF is a part of final post-processing, together with Fill Panelists Activity, where HPF concerns only non-audited nodes and FPA precisely corrects audited nodes using the site-centric data as a reference.

When HPF detects a heavy panelist (i.e. a panelist who generated on his own a significant part of the node’s Views or Time) the algorithm radically reduces all the statistics for this heavy user - by setting them to be equal to the number of the panelists Visits on this node (as a result each Visit has one View and lasts 1 second). We do not remove (filter out) the whole panelist, but only the part of the activity on a single node which happened to be excessive.

JAR (Joint Apocalypse Response) – “an umbrella term” created by Gemius for full response to the Third-Party Cookies decommissioning.

For many years internet communities have been asking for improved privacy and, as a consequence, browsers have dropped their support for Third Party Cookies (TPC). They are no longer supported by default in Safari, Microsoft Edge, and Mozilla Firefox. Google Chrome is going to terminate them by the end of the third quarter of 2024. The main goal of the JAR is to allow the continuity of gemiusAudience and gemiusDirectEffect studies in the world without TPC.

JAR consists of 2 methods:

1. Cookie Matching (CM) – cookie panel restoration,
2. Browsers Number (BN) – site-centric data restoration.

Cookie Matching (CM) – a method assessing the probability that First-Party Cookies coming from different domains, which we measure as separate ones, in fact, come from one device, and then connecting them into groups accordingly.

The probability of two FPCs coming from one device is calculated by a Machine Learning classifier, which has a very simple idea behind it:

- if two identifiers are often seen in the same place (i.e., under the same IP), at the same time, then the probability that they come from one device increases,
- if two identifiers are often seen in different places, at the same time, then the probability that they come from one device decreases.

Based on this probability, we connect FPCs together into groups called “communities”, which are equivalents of Third-Party Cookies. We try to restore all TPCs on a given market, but what interests us the most is the Gemius Cookie Panel, so we keep only communities with our survey (i.e., with known demography). CM is not limited to connecting identifiers from one identifier's space (e.g., one browser). It can connect identifiers from websites, applications, and players coming from one device, and by doing that to go even further than just restoring TPCs.

Browsers Number (BN) – a set of methods for calculating the site-centric estimate of the number of browsers (up to now called EC) for audited domains, groups, and the audited part of the Internet node, without any use of Third-Party Cookies. BN uses exclusive data available after the expected “cookie apocalypse” and it is meant to replace the current EC Global and EC Standard, which rely on TPC. Data sources used by BN, instead of TPC, are First-Party Cookies and IP addresses. Current method - EC, needs a representative sample of cookies, so-called “good cookies”. The characteristics of good cookies are extrapolated on all cookies (using the “good cookies per good PVs” ratio). BN also needs a representative group, whose characteristics would be stretched on the whole traffic, but in this case, these are representative groups of First-Party Cookies or IP addresses.

Methods within BN are designed and optimized for a few dimensions: two regarding time - day and pseudo-month (28-day month); and three regarding the object of interest - single audited domain, multiple audited domains, and audited part of an Internet node. So, altogether there are 6 combinations, thus 6 slightly different but well-justified approaches, to find the representative groups of data. Finding such representative sub-groups is the heart of BN and the most valuable Gemius' know-how.

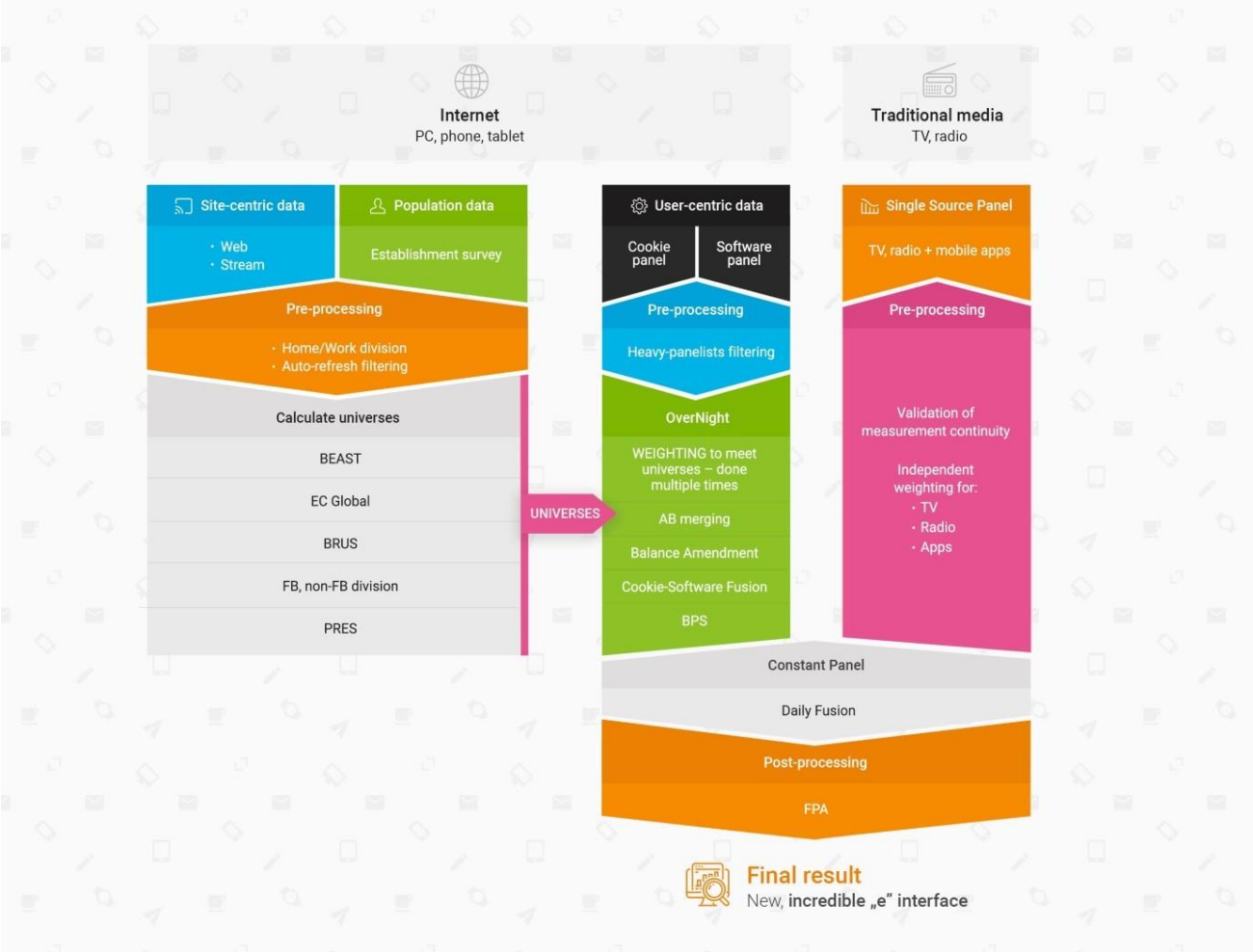
3. Overview and scheme of gemiusAudience study

gemiusAudience is a digital audience study. Our goal is to deliver the best, fastest, and most useful results for the total audience for internet nodes like websites, streams, and apps. Now, thanks to new Single Source Panel data, we also deliver cross-media results, combining the Internet with classic media like TV and Radio.

We might say that currently gemiusAudience stands on 4 pillars.

1. The first pillar is the **data sources**, obviously. We have many of them and they have different scopes, characteristics, and weak and strong points, but what is important, they all provide us with different parts of the same picture.
2. The second pillar is the **OverNight methodologies**, which stem from many years of experience and continuous improvement in fusing these different data sources into one, coherent picture.
3. The third pillar is the **Constant Panel** which, simply put, is a methodology making these OverNight results stable and consistent in time, and as a result giving us easy access to any desired custom time period.
4. The fourth pillar is the **Single Source Panel**. As the newest addition to our data sources, which brings a lot of new information and new possibilities, it deserves to be listed individually. It is our way to reach media other than the Internet (i.e., TV and Radio), completely redefining the range and meaning of the whole gemiusAudience study.

Below we present a schematic depiction of the gemiusAudience study, with visible connections and dependencies between individual parts of the methodology and the general data flow.



First, we collect all data sources, i.e., site-centric measurement data, population data, user-centric data (in the form of Cookie Panel and Software Panel), and also, if available, the **Single Source Panel which**, we strongly believe, **is the future of cross-media market research**.

Then, there are steps of pre-processing and main OverNight data processing. At this stage the majority of the methodologies developed and used by Gemius take place (described in the section “Intuition and short definitions” of this document).

After that, the calculated OverNight results are turned into a Constant Panel with the use of Daily fusion, repeated daily. Constant Panel methodology might be vast and complicated, but the main goal of it is simple: to create a stable version of the BPS panel, with low, controlled error over long time periods.

At the very end, we have post-processing in the form of the Fill Panelist Activity algorithm (and Heavy Panelists Filtration – *not yet updated on the graph above*) and presentation of the final results in the new, intuitive “E” interface (e.gemius.com).

Gemius S.A

D48 Building, Domaniewska 48

02-795 Warsaw, Poland

+48 22 390 90 90

+48 22 378 30 50

contact@gemius.com