Personal Data

Crobox

Leonard Wolters, Revision 0.7, May 2018

Table of Contents

1 Introduction

This document contains everything you need to know about Crobox and Data Processing. Specifically, we want to discuss the processing of Personal Data.

This document is structured as follows:
1. General description of Crobox Data Processing
2. High-level data flows
3. Extensive information regarding the personal data collected
4. Accountability of Artificial Intelligence and Machine Learning

1.1 Document history

0.7	2018-05-03	Updated terminology in reference to Crobox services
0.6	2017-10-17	Added subchapter Persuasion Profiling and expanded chapter 5 (AI)
0.5	2017-10-05	Added chapter about Data Processing and Data Flows
0.4	2017-09-28	Added Storage and Retention Period chapter

2 Data Processing

Crobox, the Supplier of the Services, acts as a Data Processor and processes - on explicit request - data from a Data Controller - also referred to as client - on behalf of their Data Subjects. Data Subjects are website visitors or passive individuals browsing the Internet using desktops, tablets, or any other (mobile) device.

Crobox collects *event-driven* data from Data Subjects sent by primarily web browsers[1] to build profiles - personal, page, and product-related - that encode statistics and behavior, and are solely used for marketing purposes.

Using these profiles, Crobox provides its clients a sequence of independent and isolated experiments that offer a personalized marketing experience on top of an omnichannel strategy, though often restricted to online channels only.

All experiments are part of a maturity model called the "Journey to Influence," which encapsulates a multi-year, high-level strategy offering structure, cohesion, and order to the experiments conducted.

---

[1] See https://en.wikipedia.org/wiki/Web_browser

2.1 Subject Matter

Services provided by Crobox concerning the processing of data can be broken down into two components:

1. Collecting onsite (click) data to build, manage, and maintain shopper profiles.
2. Offering (direct) marketing services including behavioral targeting (onsite), expressed by serving personal promotions on the client's online channels, e.g., their website(s).

For the purpose of above, analytics with tracking is required.

2.2 Nature and Purpose

By client's request, Crobox tracks the (click) behavior of Data Subjects, primarily denoted as website visitors, to build, manage, and maintain data profiles - both individual, page, and product related - to be used to serve personal promotions.

Data profiles are required to:

1. Offer a personalized marketing experience, primarily targeted at (direct) marketing matters.
2. Measure overall performance and filter generalizations, which might serve as input when serving personal promotions, and are required to track performance.

Subject Data that is collected, processed, and encoded into data profiles is **never** without the explicit consent of the Data Subject and is done on **explicit instruction**, and for the benefit, of the client.

Crobox's context behavior is interpreted as - though not limited to - tracking events and page views of a specific Data Subject to record whether he/she engages with a specific link, image, or product. These events can be categorized as:

1. Tracking Page Views: We track which pages (controlled by the client or Data Controller) a Data Subject visits.
2. Tracking events that denote an action including:
   a. Clicks: We track which products a Data Subject clicks, either with or without a Crobox-served promotion.
   b. Add-to-cart: We track which products are added to a shopping cart by a Data Subject.
   c. Transactions: We track if a Data Subject has successfully converted into a buyer by conducting an online transaction. Crobox detects a transaction by tracking a

page, e.g., a "thank you page." We don't store any personal information related to (financial) processing of the transaction or integrate with external Payment Providers(s).

    d. Custom: On the rare occasion, where the client doesn't offer an (online) e-commerce driven business, we might track custom events for the purpose of determining the performance of our Services. These custom events are implemented with the explicit instruction(s) of the client or Controller.

3. Personal Promotions: We offer personal promotions to Data Subjects and track whether or not a Data Subject engages with these promotions, which is required to determine our performance and adjust data profiles.

The reason why we need to collect and store this data is to adjust and maintain product and page related statistics including, but not limited to, trends, popularity, price, stock information, etc. These statistics serve as input for some of our promotions to render appropriate messages.

> Crobox will only act as the (Data) Processor within the meaning and context of the GDPR. Crobox only processes the data for which it has received the client's written instructions. Crobox will not outsource the processing of the Data, either whole or in part, to third parties unless the client has granted its prior written consent.

2.3 Shopper Profiling

The purpose of collecting data is to *optimize* the serving of personal promotions. The optimization process does not rely completely on the personal data of the Data Subject; it relies on the *full context,* which also includes (aggregated) information about the products and the pages the Data Subject is subject to.
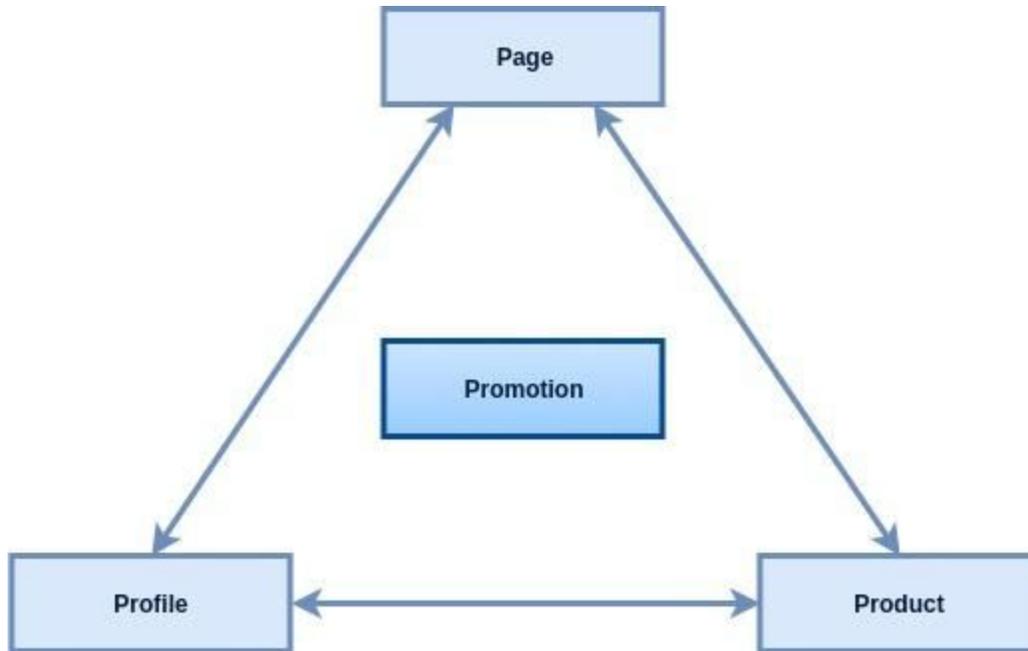
To simplify, serving a promotion is based on evaluating the current Page, Product, and Profile. Based on these three subjects, the best promotion is determined and served to the Data Subject.

The reasons for combining these subjects include:

1. Data Subjects alone often don't have enough data points to come to legitimate predictions (i.e., the "cold start" problem).
2. Some products (or pages) form robust combinations with particular promotions - no matter the impact of the Profile.
3. By taking into account several distinct "forces" or sources, the Data Subject experiences more diversity (in promotions).
4. Machine Learning improves with the number of attributes, e.g., single pieces of information that make up the Visitor Context.

It is important to know that we <u>do not base personalization only on the data of the Data Subject</u>. We also use the aggregated information included in the full Visitor Context. As a result, the impact of the Data Subject is limited and/or restricted.

2.4 Data Subject

2.4.1 Matters

The Data Subject Matter of the processing of data comprises the following data types:

1. Customer History (although only Online)
2. Online Click Data

For a better understanding of what we exactly collect, please consult 2.2. Nature and Purpose.

2.4.2 Categories

The categories of Data Subjects comprise of:

1. Customers
2. Potential Customers
3. Website / Online Visitors

2.4.3 Unique Identifier (UUID)

Crobox uses a single piece of information required to identify a Data Subject, a generated unique identifier (UUID), consisting of a unique sequence of 40 randomly picked characters and numbers, that is stored in a Cookie residing in the Data Subject's browser. This UUID is used to identify, connect, and appoint personal Crobox profiles with the Data Subject.

We do not store any other personal contact information, including email, physical and online addresses, names, etc.

2.4.4 Personal & Sensitive Data

Besides knowing what data Crobox collects, it is also important to know what information and (personal) data Crobox doesn't collect. Crobox limits collecting data to what is required to render its Services.

Crobox will **never collect, store, process, generalize, or construct** the following Sensitive or Personal Data:

1. Personal contact details including name, address, phone numbers, or email addresses
2. Online identity information including - but not limited to - username/password combinations
3. Data revealing racial or ethnic origin
4. Data concerning political opinions
5. Data concerning religious or philosophical beliefs
6. Genetic or biometric data
7. Data concerning (personal) health
8. Data concerning a natural person's sex life or sexual orientation

9. Data concerning the economic and financial situation of the Data Subject. Including any financial related information, e.g., bank account numbers, personal pin codes, credit card numbers
10. Data about Data Subject's performance at work
11. Data about movements

Crobox does store a location - in restricted form - of the Data Subject which is based on the IP address encapsulated in online traffic. When possible, the IP address is translated into a city/country/region triple which is stored within the Visitor Context.

We do not store/keep the IP address nor the latitude/longitude combination.

3 Data Flow & Collections

Below you'll find a high-level overview of the data flow of Crobox's Services.
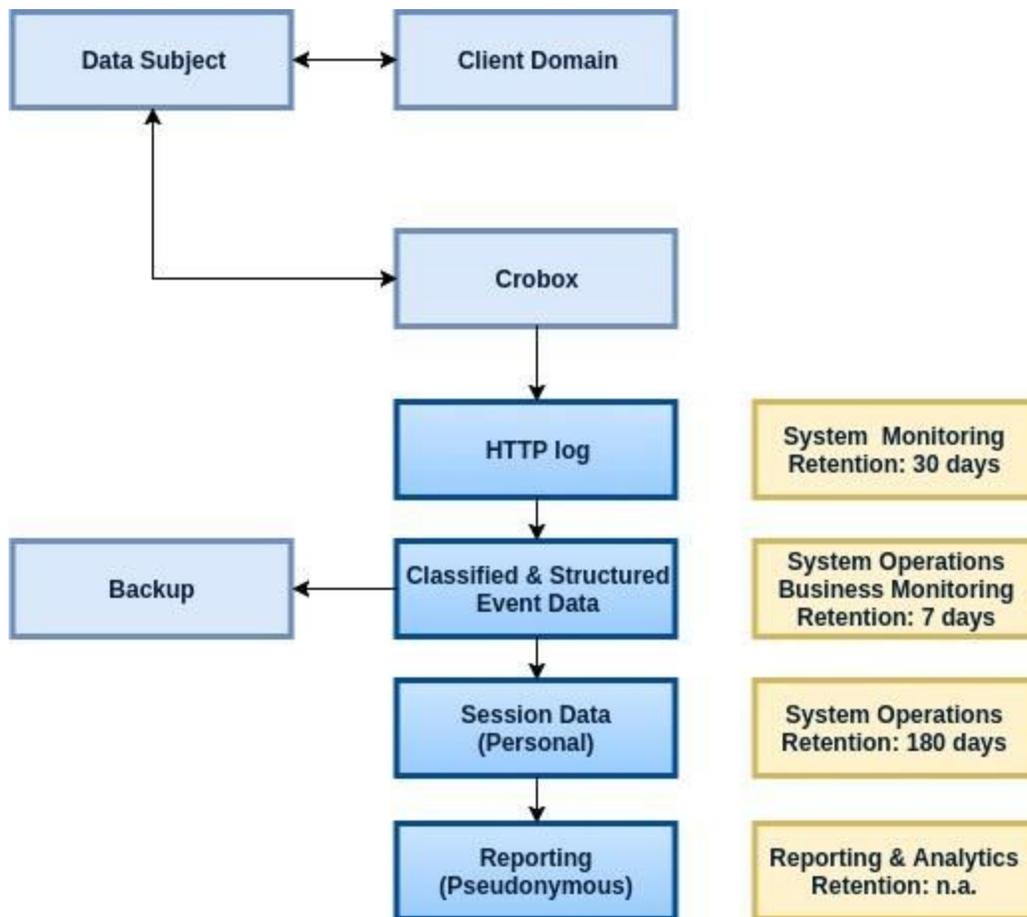
Some more information about our data collections (i.e., the dark blue boxes in the center):

1. **HTTP Log**: HTTP log is considered raw log information that comes directly from our HTTP servers. This includes personal data such as IP addresses but is only used for system monitoring. This data is transformed into Event Data but is fully removed afterward. Our system monitoring tools currently have a retention period of 30 days, after this period all information is automatically deleted.
2. **Event Data**: The HTTP log, mostly consisting of HTTP requests, is transposed into classified and structured Event Data. Event Data has validated data that belongs to a Data Subject and a client. Event Data is considered immutable: it represents a single action or request that took place at a current moment in time. This information is processed into Session Data that contains aggregated information. Event Data forms the heart of our platform, it is considered the source of truth and, therefore, the only data collection being (permanently) backed up. Besides being backed up, Event Data is also closely monitored (for business monitoring purposes) and has a retention period of 7 days.
3. **Session Data**: Session Data holds all Event Data belonging to a single session and also adds some aggregated information/intelligence to it. It has a retention rate of 180 days by default.
4. **Reporting**: Reporting data only contains aggregated information and as a result, doesn't contain any personal data. This data is used for analytics. Reporting Data exists for the duration of a Service Agreement with Crobox.

3.1 Retention Period Personal Data

After 180 days of inactivity, i.e., without the Personal Data belonging to a specific Data Subject being updated, it will be automatically deleted. Deletion includes full removal of all **Session Data** - i.e., the Visitor Context - associated with the specific Data Subject from our primary record of store.

In short:

1. All personal data will be deleted when the Service Agreement between client and Crobox ends,
2. All personal data can be   deleted upon request, i.e., the "Right-to-be-forgotten",
3. The cookie used to identify a user has an expiration date of 180 days after the last session.

> All personal data, which is represented as a sequence of immutable events, is immediately aggregated into general and accumulated statistics (i.e., Reporting Data), thereby losing its personal characteristics.

3.2 Backup Files

To provide a fluent service delivery, including disaster recovery, Crobox periodically backs up system data (i.e., *Classified and Structured Event Data*). These backup files include Personal Data.

Backup data is separated per client, pseudonymized, and encrypted using a proprietary binary format. Deleting personal data (i.e., "The Right to be Forgotten") does not affect these backup files, since altering these files (taken into account backward propagation) is too costly and too cumbersome. Besides, they are encrypted and stored in an inaccessible place.

Also, when Data Subjects indicated to be forgotten, Crobox ensures their personal data is not restored when performing (disaster) recoveries.

4 Personal Data

The graph below represents a schematic overview of all the Personal Data Objects collected by Crobox.
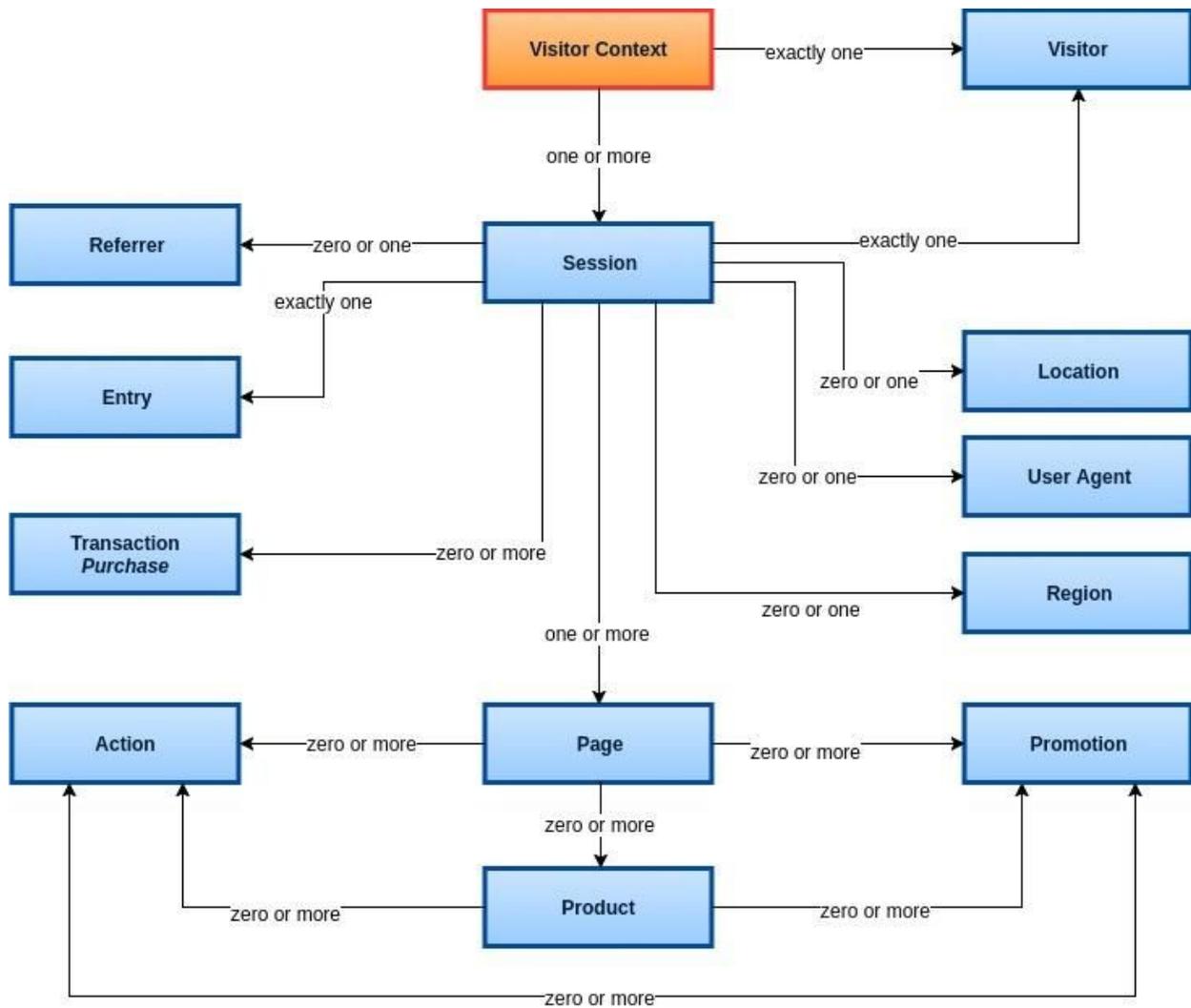
Reminder: the purpose of collecting data is to optimize the serving of personal promotions (2.3 Shopper Profiling). Please see 5 Artificial Intelligence & Machine Learning for more information about how this data is used and processed.

4.1 The Visitor Context

The Visitor Context is the container that holds all (personal) data belonging to the Data Subjects and is collected, managed, and constructed by Crobox. This object is present in the Session Data as described in 3 Data Flow & Collections. It holds the following data objects:

1. Visitor: Each context must always have a visitor attached to it. The visitor contains key information about the Data Subject.
2. At least one or more session(s): A session holds information about the Data Subject within a single website visit.

4.2 Visitor

The visitor holds relevant information about the Data Subject, i.e., the owner of the context. The visitor stores a unique identifier (i.e., UUID) generated by Crobox representing the Data Subject. This UUID is also stored in the cookie and is, thus, used to retrieve and match appropriate (personal) data.

4.3 Session

A session holds all relevant information belonging to a single visit of a web page (or mobile app). Usually a session represents a "full" visit and is comparable to a visit in a physical shop, meaning that it stops whenever you leave the (web)shop. For detecting when someone leaves a webpage, we rely on Browser functionality. Next to that, we automatically end a session when we haven't received a piece of new information within thirty minutes.

A new session can be started whenever the previous one has ended.

The session itself contains the following important data, which will be discussed in separate paragraphs:

1. **Referrer**: The last URL the user visited before arriving at the client's website.
2. **Entry**: The page (denoted by URL) that the visitor landed on when navigating to the client's website.
3. **Location**: Based on the IP address of the visitor, we map this to Geo-Location, if possible.
4. **User Agent**: The browser the visitor is using.
5. **Region**: The region of client's website the visitor is currently visiting, which holds information about country, language, and currency.
6. **Transactions**: The number of products the visitor successfully bought on the client's website.
7. **Page**: Holds information about the single page that the visitor visits during their session on client's website.

4.4 Referrer

A referer holds information about the last website visited *before* entering the client's domain. When available, we map the URL of this website with the following data points:

1. **Medium**: What was the medium or "type of business" of the Referrer? We currently support the following values: Email, Search, Social Media, Affiliate, and Internal.
2. **Name**: The qualified name of the referrer, such as Google, Facebook, Nu.nl, etc.

4.5 Entry

The entry object represents the place where the visitor lands. This information might come in handy since it reveals if people are entering at the front door (i.e., the "home" page) or directly on a content/product page.

The Entry object is currently represented by a URL.

4.6 Transaction

We track the following information per transaction when available:

1. Price per purchased product
2. The unique order ID, which is often required to validate our tracked transactions with the orders of client's backend system(s)

4.7 Location

A location is determined by the IP address of a visitor (during the session). We use external services to map the IP address to a location consisting of:

1. **City**: The city determined based on given IP address
2. **Country**: The country determined based on given IP address
3. **Region**: The region determined based on given IP address (Europe, Asia, etc.)
4. **Population**: Based on the detected city, if available

It is often not possible to determine or extract this information by someone's IP address.

It's important to note that after this lookup - either when it succeeds or fails - we **discard** the IP address.

4.8 User Agent

The User Agent holds information about the Browser used by the visitor for given session. It holds basic information such as:

1. **Browser**: For example Firefox, Chrome, or Safari
2. **Version**: The version of the browser used
3. **Device Type**: The kind of device the visitor is using during his session, e.g., mobile, personal computer, or tablet.
4. **Operation System**: The operating system the visitor is using.

Keep in mind that a User Agent cannot always be detected.

4.9 Region

A single website can target multiple countries, multiple languages, and
can support multiple currencies in some scenarios. A unique combination of these three values
is called a region. Only one region can be active at a time for an active session.

A region contains:

1. **Country**: The country of the website currently visited by the visitor.
2. **Language**: The language currently selected or preferred by the visitor.
3. **Currency:** The selected currency of client's website.

4.10 Page

A page contains relevant information used by our platform, of which the most important tracked
items are:

- Products listed on the page
- Promotions present
- Actions performed (e.g., clicks on a button)

A page is a unique rendering of a URL that resides on the website of the client. For each page,
we detect what kind of page it is (e.g., Home Page, Product Listing Page, Product Detail Page,
etc.), which products are listed, and which promotions (served by Crobox) are shown.

4.11 Product

The Product Data entity represents a product on the client's website and is denoted as SKU
(store keeping unit) or a unique product ID. If possible, we scrape other relevant information,
such as the current price used or the category where the product belongs.

If available, a client might provide us a Product Feed that contains additional product
information, such as stock, color, material, etc. We try to combine this feed data with product
data retrieved from the website (by matching the product IDs from both sources) to optimize
promotions.

4.12 Promotion

A promotion is a unique serving of a personalized marketing message from Crobox's platform. A
promotion holds relevant information such as:

1. **Message ID**: Type of message, each experiment can have several different messages representing different content and/or strategies.
2. **Promotion Type**: Promotion used, e.g., Product Tags, Smart Notifications, Exit Intent.
3. **Principle Type:** The underlying principle used for each promotion.

We track promotions for performance matters. Crobox evaluates the promotions belonging to certain products or visitors and the corresponding actions to track behavior and build messaging strategies.

Some side notes:

- Promotions do not necessarily need to be related to a product. For example, an Exit Intent is not coupled with a particular product.
- Promotions might have *multiple* actions defined.

4.13 Action

An action denotes a single action the user performs on the client's website, like a click. Crobox only tracks product and promotion related actions. Out-of-the-box, we support the following:

1. **Product Clicks**: The products that are clicked on the Product Detail Page
2. **Add-to-carts**: Products added to the cart
3. **Transactions**: Successful orders that can be tracked based on monitoring success pages, For example, when a Payment Service Provider (PSP) redirects back to the client's domain to show a successful transaction.

In some cases, we might support custom actions, which are only done on explicit request and written approval of a client. For example, custom actions would be "Add to Wishlist" or "Share on Facebook."

5 Artificial Intelligence & Machine Learning

Crobox uses Machine Learning (ML) and Artificial Intelligence (AI) to optimize serving personal promotions. Currently, we apply machine learning at the following levels:

1. Product Targeting
2. Promotion Targeting
3. Experiment Targeting

5.1 Product Targeting

One core element of our ML is to predict which products are most suitable to enrich with promotions. Product targeting is done at an experiment level. Products are targeted per experiment to create the best match for the Visitor Context and are in the best interest of that particular experiment. As a logical consequence of this, different products can be targeted given the visitor Context of the Data Subject.

Product targeting only takes into account which products a Data Subject has seen or interacted with in a given experiment. Based on this selection of products, similarities with other Data Subjects can be calculated and used to recommend or target products that are shared by the closest peers but not yet seen by given Data Subject.

The techniques used all belong to the field of Collaborative Filtering (CF).

5.2 Promotion Targeting

Once products have been targeted, the next step is to determine the best promotion that fits selected products and matches the visitor Context (i.e., *promotion targeting)*.

The essence of promotion targeting is in the heart of "contextual multi-armed bandits," see https://en.wikipedia.org/wiki/Multi-armed_bandit. We combine principles from this theory with Random Forests (https://en.wikipedia.org/wiki/Random_forest) to optimize serving promotions.

In short:
The Visitor Context consists of many small single pieces of information called features[2]. Some examples of features: browser used, number of pages visited, entry page, product category, etc. Since we don't know beforehand *which* of these features (or selection thereof) is most "discriminative." Machine Learning keeps track of all different (often random) combinations of feature values and keeps track of which do and do not promotions work. Based on all these combinations and their statistics, random samples are drawn using the feature values found in the visitor Context and, based on all these samples, the best promotion is determined.

Given the nature of our algorithms, techniques, and methodologies - which are based on taking (random) samples - it is extremely hard to justify exactly - or deterministically - which and why a particular promotion has been selected given the context of a Data Subject.

A promotion is selected at a particular moment in time using the current product, page, and profile statistics (see 2.3 Shopper Profiling), which all continuously change.

---

[2] These features are either attributes of the objects found in 4 Personal Data or are closely related to them.

5.3 Experiment Targeting (2018)

The final segment where Machine Learning is applied is to the selection of experiments. Crobox offers its clients a maturity model entitled the "Journey of Influence" that consists of a collection of individual and independent psychology-based experiments. Some of these experiments run in parallel, others don't. Therefore, given a visitor Context, an optimal path (i.e., selection of experiments) can be determined.