Practical Data Dictionary

Mainly_for_online businesses

Created by
Tomi_Mester

Published on data36.com

Why_is_this important?

When a company begins to use data, they usually read a bunch of articles and books on the subject. In good cases, they hire 1-2-3 data analysts and set up a data infrastructure and/or a data strategy. Then slowly everyone starts to use the resulting data in the company and an awesome data-driven organization is born. Hooray!

But along the way there will be some disorder caused by the use of materials pulled from various sources, and people's different know-how. Because Data Science is not a written in stone kind of science, it's not uncommon for the same concept to be known under another name in different places. What's even more crazy is that this is true the other way around as well: the same word can be used for many different concepts as well.

Working on different projects I realized, this issue became increasingly problematic. For this reason, I decided to create a dictionary which unifies such data expressions and places them within a clear framework. The main points were:

- consistency
- simplicity, so not having to memorize 800 different types of users (created 8 categories for activity, and 5 for payment)
- expressions for particular things should resemble each other as little as possible (not to have 3 different but similar-sounding categories, like Active User, Activated User, Re-activated user, etc.)

This is how **Practical Data Dictionary** came about, which I will open-source as maybe others have also experienced these kinds of issues. I advice this booklet so everyone within the organization speaks the same language, and to communicate about data quickly without any misunderstanding.

Content

Hi, I'm Tomi Mester, I am the editor of data36.com blog since 2014. (Before that, I was a Data Analyst at Prezi.com.)
My main goal with Data36 is to spread data-driven thinking in Europe (and all over the world) to help as many businesses become better and better as possible.

We could have met before as I also give presentations sometimes in conferences on this topic, like e.g. TEDxYouth, the Barcelona E-commerce Summit, Business Intelligence Forum, etc...

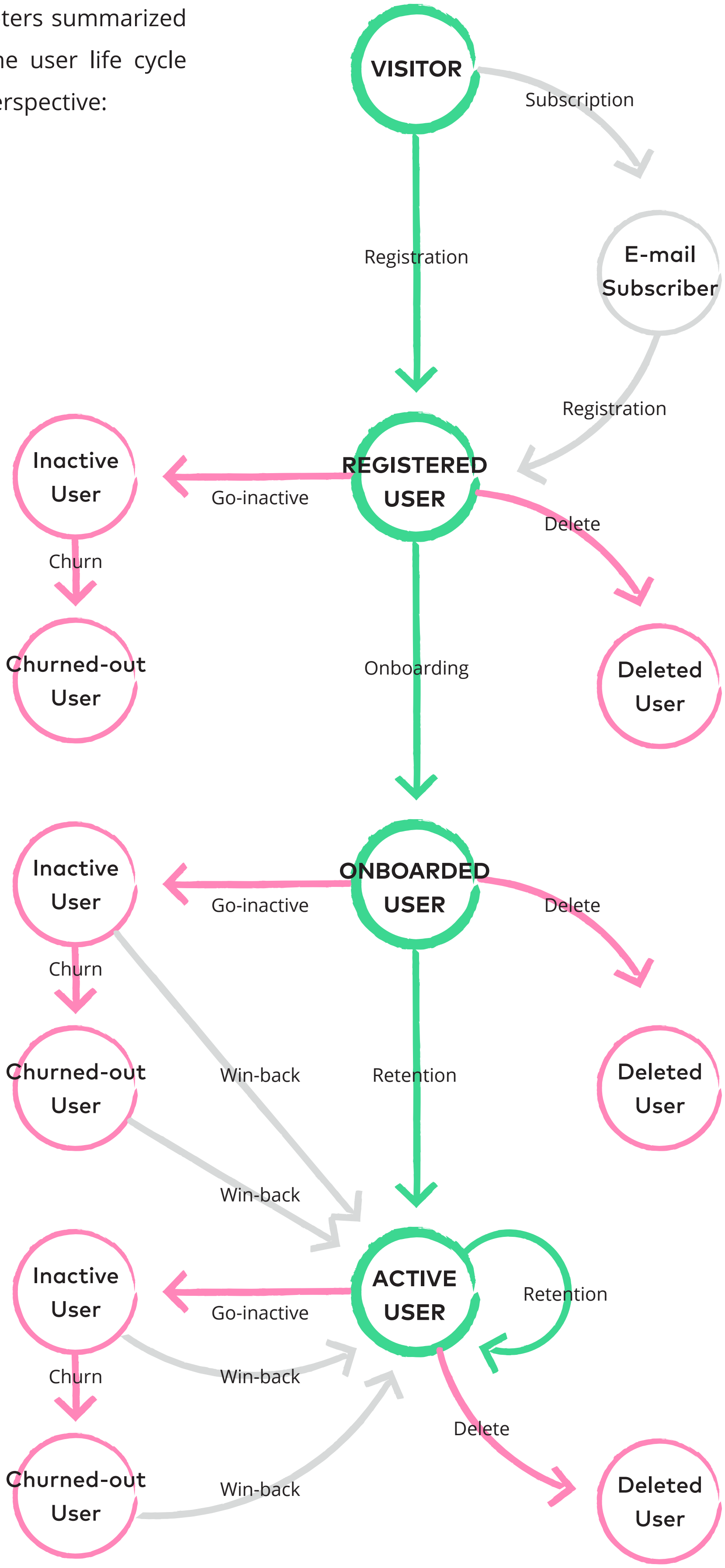
For more info, click below:
My LinkedIn profile: https://se.linkedin.com/in/tomimester
My E-mail address: tomi@data36.com
Follow me on Twitter: https://twitter.com/data36_com

A little about me

Chapter_01

Activity-related events

The first two chapters summarized in a diagram – the user life cycle from an activity perspective:



Visit

When someone visits our webpage.

E-mail subscription

E-mail Subscription: When someone visits our webpage and provides their e-mail address – but may not necessarily create a user account. This is most commonly signing up for the newsletter.

Registration

When someone visits our webpage and creates a user account, and provides at least one unique identifier (e-mail address, FB account, stuff like that).

Onboarding

(Usually the process which takes place right after Registration) during which the Registered User goes through the key steps which make up the basis of our product. It's during the Onboarding that the User becomes familiar with the main values of our product (e.g. has added 5 friends on the social media app and wrote at least one post; created and sent the first invoice in an invoice issuing software, etc…)

You need to define your Onboarding process, and it's worthwhile to create it in a way to enable the user to see the value of your product by the end of it, so they will use your product or service again and again. (E.g. writes newer and newer posts, sends newer and newer invoices, etc…)

Note: It can happen that Onboarding has an "ideal time-frame", but I think this is pointless, because if someone does not go through the Onboarding, they will become an Inactive User, then a Churned-out User anyway.

Retention

Keeping the users - an Active User will continue using our product, they will use our product/service again and again and will become/remain an Active User.

Note: If the user logged into her user account, it does not necessarily mean that she used our product as well. You'd actually be surprised to see the ratio of the logged-in-but-did-nothing-else user ratio on many product… It is worthwhile to link activity identification to the end of the Onboarding process: it's often suggested to make it the very end (e.g. with an invoice issuing software: they logged into their account --» we don't consider this activity; they sent another invoice --» this is considered activity).

Go-Inactive

When a user does not use our product/service for a given time period (or above that).

Churn

When an Inactive User does not use our product/service for a given time period (or above that).

Win-back

When an Inactive User or a Churned-out User becomes an Active User again.

Delete

When a User deletes themselves or asks us to delete them from our system.

Chapter_02

User-types from an activity perspective

Visitor
Someone who visits the website, a potential Registered User – but not necessarily one.

E-mail Subscriber
A visitor who provides their email address.

Registered User, in short: User
The kind of Visitor who registers, so provides their email address, their Facebook account or any kind of unique identifier, for which we create a user account.

Onboarded User
A User who has gone through the so-called Onboarding-process.

Active User
This is a changing status. The kind of user who uses our product in a specific time-frame marked by us (e.g. a given month, given week, given day or given hour).
Note: Again! If the user logged into her user account, it does not necessarily mean that she used our product as well. You'd actually be surprised to see the ratio of the logged-in-but-did-nothing-else user ratio on many product... It is worthwhile to link activity identification to the end of the Onboarding process: it's often suggested to make it the very end (e.g. with an invoice issuing software: they logged into their account --» we don't consider this activity; they sent another invoice --» this is considered activity).

Inactive User
This is a changing status. The kind of User who does not use our product for a specific time-frame marked by us (e.g. a given month, week, day or hour).

Churned-out User
This is a changing status. The kind of User who has not used our product for a specified, lengthy time-frame marked by us (e.g. the past 3 months, past 1 year, etc.).

Deleted User
The kind of User who we deleted from our system or who has deleted themselves.

Note1: If you check the process diagram again, it will be clear that the E-mail Subscriber, Registered User and Onboarded User status' are one-time status'. The main goal is to push our Users through these – as many as possible – and to keep them as Active Users for as long as possible. This will not work with everyone of course. From this it follows that there will be relatively low Users in the E-mail Subscriber, Registered User and Onboarded User status. Most of the Users will be coming and going between the Active/Inactive/Churned-out status'.
However, it's still worthwhile to have the E-mail Subscriber/Registered/Onboarded categories segmented as these Users are very fresh and curious. Due to this, they are „sensitive" about many things, thus they are easy to handle, ideal Users for you.

Supplement to Chapter_02

Derivative user types from an activity perspective

During User research, we aren't only interested in what phase they are in now (Onboarded, Active, etc.), but also in what phase they were in before. It makes a difference whether an Inactive User – prior to Inactive status – only registered and did not try the product yet (was a Registered User), or he/she tried the product, but only once (he/she was an Onboarded User), or he/she used it often (was an Active User). It's sometimes advised to segment the users from each other from this perspective as well.

INACTIVE USER SEGMENT

Registered-then-Inactive User

The User who after Registration immediately became an Inactive User.

Comment: Another coined term is Dead-On-Arrival.

Onboarding-then-Inactive User

The User who after Onboarding immediately became an Inactive User.

Active-then-Inactive User

The User who was Active, but then became Inactive.

ACTIVE USER SEGMENT

Onboarded-then-Active User

The User who went through the Onboarding process and stayed an Active User.

Active-then-Active User

The User who was an Active User and stayed an Active User.

Inactive-then-Active User

The User who returned after Inactive User status (Win-back) and then became an Active User.

Churned-then-Active User

The User who returned after Churning status (Win-back) and then became an Active User.

Note1: It could be interesting to broaden these groups based on our own preferences. E.g. 5*Active User (the User who was an Active User 5 weeks straight), etc...

Note2: At the same time, it's not worthwhile to create too many subcategories either as it's easy to lose focus if we concentrate on many segments.

Note3: Since we touched on the topic of focus! It's a basic question of strategy on which of the above categories (8 + 3 + 3 + your own subcategories = 14+) we concentrate on. A lot of literature exists on why it's better to pay attention to the Registered Users rather than the Inactive Users, or why Win-back is more valuable than Retention. These are interesting reads... BUT! Your product, your strategy and your Users will determine who you will focus on – for this you need to analyze your data, and not follow other people's advice. Check it out and decide what's important for you and with measurements identify what you need to place in the center to achieve this.

Another
Supplement to Chapter_02

User groups on a time basis

The above User groups are more easily manageable if you divide them into groups on a time basis (e.g. Daily Active Users). Based on our personal experience, it's practical if these belong to not relatively but absolutely determined time periods. So we are not watching those who were Active Users in the past 24 hours (as this is a constantly changing group), but those who e.g. were Active Users between 2016-01-01- 00:00 and 24:00 (as this is a fixed group, once 2016-01-01 24:00 has passed, then the distribution of the group does not change).

These groups also need to be generated by you based on your needs, but here are some examples:

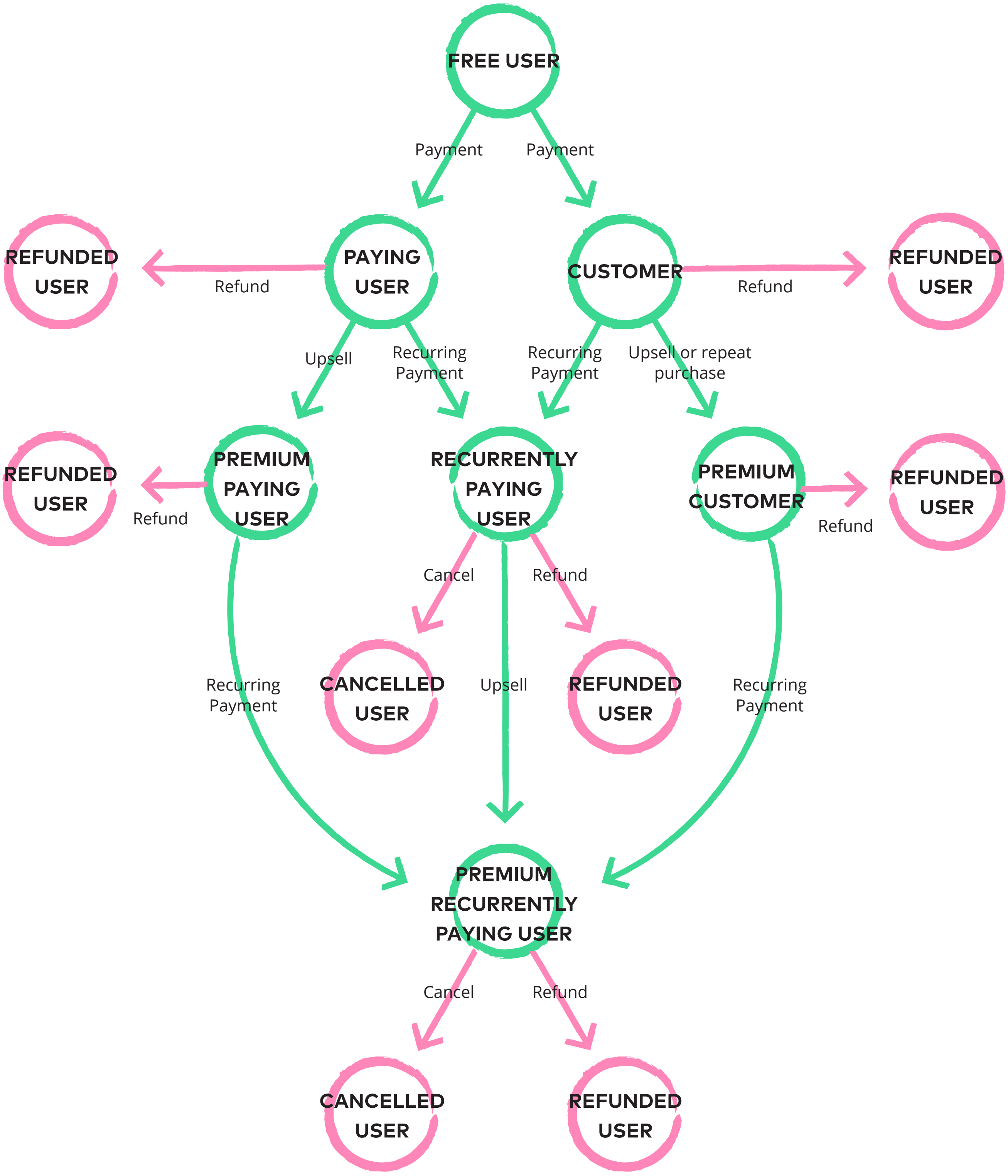
- Daily Active Users (e.g. the number of Active Users on 2016-01-01 is: 352)
- Weekly Onboarded Users (e.g. the number of Onboarded Users on W1 of 2016 is: 1.860)
- Yearly Churned-Out Users (e.g. Churned-out users in 2015 is: 21.512)
- etc, etc...

Chapter_03

Payment-related events

Another summary diagram – this time with reference to chapter 3 and 4 – the user life-cycle from a payment perspective.

Note1: Payment models can be highly varied, so don't be surprised if the below diagram is not relevant entirely to your business, but just to a small part of it.



Payment A payment event, transaction. The purchased thing can be a specific product (e.g. a pair of shoes) or a service (e.g. a hosting service).

Refund Returning a payment. When the Customer/User asks for their money back (and receives it).

Comment: Interestingly enough, the Refunded Users are usually a very satisfied group.

Recurring Payment Regular payment. Most common with services, but it can happen with products, too (e.g. a magazine subscription).

Cancel Cancellation of a regular subscription. Does not necessarily mean a Refund.

Upsell Selling a Customer or Paying User a more expensive product/service.

Repeat Purchase Similar to a Recurring Payment. Selling a Customer a new or given product again.

Note2: The above model and captions are too forced when it comes to Ad-click models. In those cases we are only talking about Visitors or Users, or maybe Ad-clicks, but not payments really.

Chapter_04

User-types from a payment perspective

Free User

The kind of User who has registered, may be using our product but has not yet made payment to us.

Customer

A „shopper" who has purchased at least one product from us. Not the same as a Paying User!

Paying User

The kind of User who has paid to use our service for a given time period (e.g. a premium or other payable function of our product). Not the same as a Customer!

Note: The main difference between a Paying User and a Customer is that a Paying User pays for a service which is mostly for a given time period (and can be renewed), whilst a Customer pays for a specific product once and can use it for an endless period. E.g. in this wording, if someone buys a "boxed" Microsoft Office 2015, then she is a Customer, but who subscribes to Microsoft 365 and uses the Office softwares as a monthly payable service package is a Paying User.

Refunded User

A User who for some reason asked for their money back (and received it). (E.g. she did not like the purchased shoes and sent it back; or she did not like the software she subscribed for.)

Cancelled User

The kind of User who was a Recurrently Paying User, but in the end cancelled their subscription. (But did not necessarily ask for a refund).

Supplement

Additional user type subcategories from a payment perspective

Premium Customer

A special Customer group who spend above a specific value (through an Upsell or Repeat Purchase).

Recurrently Paying User

A Paying User who regularly subscribes to a given service (in exceptional cases to a product – e.g. a magazine subscription).

Premium Paying User

A special Paying User group who spend above a specific value (through an Upsell).

Premium Recurrently Paying User

A User who regularly subscribes and spends above a specific value for a given service (in exceptional cases for a product – e.g. a magazine subscription).

Chapter_05

Summarizing what has been said

I collected into a table all the different User types based on activity and payment. For simplicity, I used 5 main categories for payment.

It's clear that this way many categories are created – if we remove the a priori impossible categories (like e.g. the paying, but non-registering user), we still have 58 groups.

This can be further expanded with your own categories. Within a large organization, it's possible for each group to have and implement its own marketing and/or product development strategy, but if this task is performed by a few people, then it's very important to find the focus. As I mentioned above, what you concentrate on should not depend on what stuff you picked up on the Internet, but more so based on the below data:

- Which group has the most people
- Which group is the most problematic for you
- Which group carries the largest potential for you

Note: The User interviews and Usability Tests can also be helpful with this data!

		FREE	CUSTOMER	PAYING	REFUNDED	CANCELLED
VISITOR				✗		
E-MAIL SUBSCRIBER				✗		
REGISTERED USER					✗	✗
ONBOARDED USER						
ACTIVE USER	ONBOARDED_THEN ACTIVE_USER					
	ACTICE_THEN ACTIVE_USER					
	INACTIVE_THEN ACTIVE_USER					
	CHURNED_THEN ACTIVE_USER					
INACTIVE USER	REGISTERED_THEN INACTIVE_USER					
	ONBOARDED_THEN INACTIVE_USER					
	ACTIVE_THEN INACTIVE_USER					
CHURNED_OUT USER						
DELETED USER		✗	✗	✗		

Chapter_06

Analytics, metrics, KPI-s

Note:

In this chapter, I was not working towards fullness. I'm going to reveal the most often used metrics – for a kind of inspiration. The aim in this part is to understand the "logic" and the exploration of problematic cases.

Chapter_06A

Rates related to events and payment

"X"-Day-Retention

This is the maximum time-frame within which an Active User needs to return in order to stay an Active User and not to become an Inactive User.

Note: The value of "X" is of key importance, yet it is a very difficult value to define. 4 principles help with the definition. The first principle is the "own-expectations" principle: we define how often we expect users to return based on our main functions. (E.g. with a news app we can expect daily frequency - 1-Day-Retention -, whilst with a travel-booking product, it can be up to 6 months – 6-Month-Retention.) The second principle is the data-centric principle: it's worthwhile to check the frequency of return based on our current data. The third is the "asap-return" principle: it is easier to measure and it's a better goal if your users come back as often as possible. For this reason, if you are unsure of whether to make the goal 3 or 4 days, pick 3. The fourth is the others-know-already principle: look for benchmarks in your own market. I dive more deep into this topic here: http://data36.com/measuring-retention/

Retention %

The ((Active User)/(Registered User)) rate within a given cohort (cohort: see below or in the above article). As we know, an Active User is someone who uses our product again and again within the X-Day Retention time-frame.

Leave %

The ((Inactive User)/(Registered User)) rate within a given cohort. Similarly to the previous point: An Inactive User is someone who does not use our product within the X-Day Retention time-frame.

"Y"-Day-Churn

The maximum time-frame within which an Inactive User needs to return to not become a Churned-out User. The "Y" value is usually a value not too far away from "X". (e.g. if 1 week is X, then one month is Y).

Churn %

The ((Churned-out User)/(Registered User)) rate within a cohort.

Win-back %

The rate of those within a cohort, who went from Inactive-then-active OR from Churned-then-active, comparing to the number of Churned-out Users AND Inactive Users who were targeted by the given Win-back campaign.

Note: It is more visible with this metric that these numbers cannot necessarily be standardized. A lot depends on what the strategy or goal is in a given campaign.

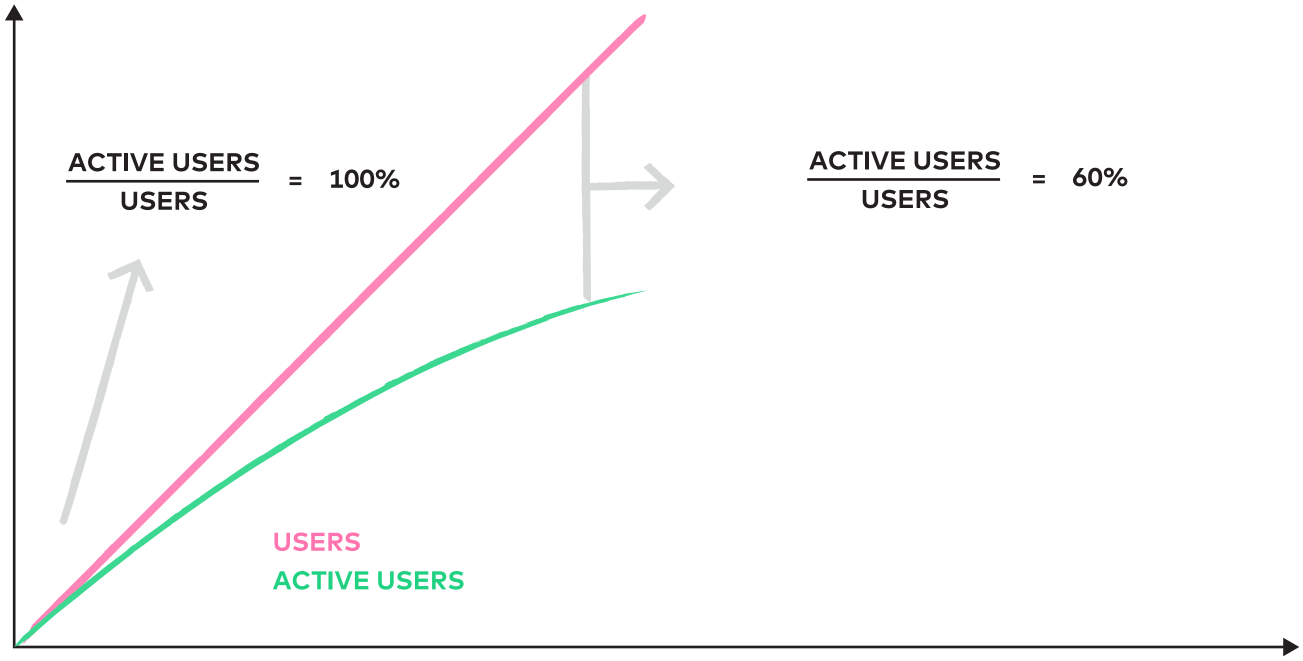
Visit-to-Registration %

The ratio of ((Registered User)/(Visitor)) on a given day (or week or month).

X-to-Y %

Based on the above examples, any ratio between two statuses' can be calculated.

Comment: Be smart in choosing which time intervals you examine! Again! If you are not checking out cohorts, you can easily mislead yourself (e.g. the (Daily Active Users) / (All Users) ratio will inevitably and continuously decrease in time. During the first few days of the product launch, most Users will be Active Users. Later, as more and more Users Churn, this ratio will constantly shift. This is natural, but because of this, an incorrectly defined ratio will not be informative at all).



Conversion %

Although this is a common expression, we don't use it often with complex products as it is too general. Conversion can be the performance of an advertisement, a purchase, a registration. Anything. It's difficult to use it in a unified way within a company.

Revenue

The generated revenue of a company for a given period. It does not necessarily show profitability, since it does not include costs. Yet in most cases, we use this as a financial KPI, as it is easily measurable.

Note1: In more complex analysis', we can actually calculate profit as well. In this case, we deduct the costs from the Revenue. The difficulty of this is that it's impossible to weigh the associated costs per product or service for an entire company, like e.g. a PR-campaign or hiring a new Head of Technology to a company.

Note2: Revenue is not just calculated on a company level, it can be done for subcategories or per product, too! See the „Segmentation" and „Case Studies" part below.

Repeat Purchase %

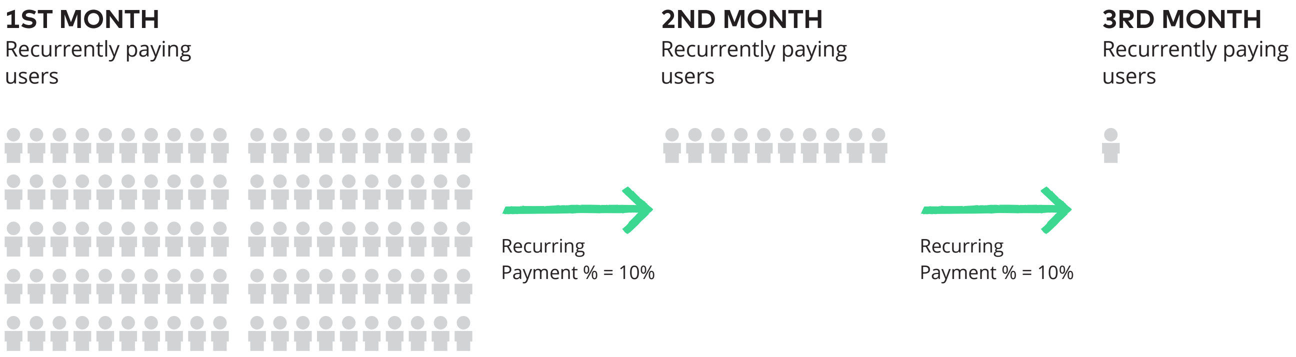
It gives the probability of a repeat purchase from a customer (provided you have what to sell).

Note: For simplicity, I usually put Cross-Sell into this category as well, so when we sell a product with another product. (e.g. movie tickets and coke)

Recurring Payment

(Similarly to the % of a Repeat Purchase) gives the probability of a Paying User to keep paying for our service. (In certain business models Recurring Payment can be automatic.)

For example, if we have a software with monthly, automatically renewed subscriptions, but on average 90% of users Cancel their subscription, then the Recurring Payment %=10%. Thus out of 100 users 10 will pay for the second month as well and out of the 10 users 1 will pay for the third month too. (Of course, it's really simplified).



Lifetime value (LTV)

Gives the average generated Revenue value of a User during his/her entire lifecycle (so up until he/she is an Active User). This value is incredibly useful for the calculation of profitability – and within that, the calculation of allowed costs. To highlight the most basic of all: it makes it simple to calculate if it's worth our while to spend „X" on a given advertisement which brings „Y" number „Z" Lifetime Value Users.

Note: On paper if X < Y * Z and we have no further costs, then it's worth it. In reality, out of (Y * Z) you need to deduct other costs and the planned profit as well.

The problem is that LTV in 99% of cases cannot really be defined, as even a Churned-out User can come back after 2 years through some miracle – and can start generating Revenue out of nowhere.

The right method depends on the business model. You can find a lot of descriptions on how to "calculate lifetime value" on the Internet. It's worthwhile reading through these carefully, handling them with criticism and checking whether they are the right fit for your business. (E.g. If you google it, I would not advise to use the first hit found on the Kissmetrics blog.) Once you have found a fitting LTV calculation method, verify if the results are realistic with a quick calculation. If yes, you're good.

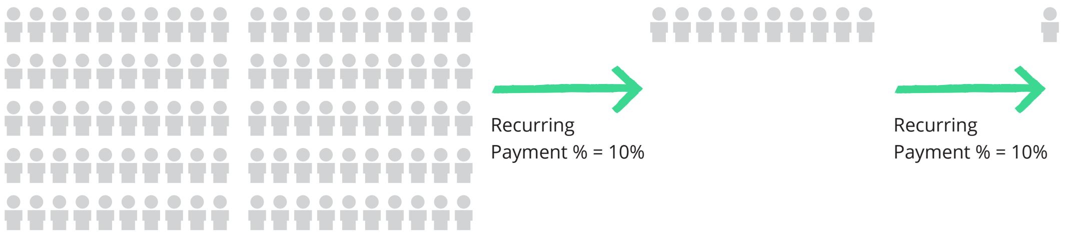
I'll show you another relatively good and simple model, which uses the Average Revenue per User (ARPU) value and the Repeat Purchase % (RP%) based on the below formula:

$$\mathrm{LTV} = \mathrm{ARPU} * (1 + (\mathrm{RP}\%) + (\mathrm{RP}\%)^2 + (\mathrm{RP}\%)^3 + (\mathrm{RP}\%)^4 + (\mathrm{RP}\%)^5 + (\mathrm{RP}\%)^6 \ldots)$$

So:
ARPU = 100$
RP% = 10%
then:

100 $ * (1 + 0.1 + 0.01 + 0.001 + 0.0001...) = 111.111 $ is the Lifetime Value

Note: In this formula, we are underestimating the LTV. When calculating the LTV, I would advise underestimating – if we are thinking in terms of money, it's better to be pleasantly surprised rather than disappointed!



1ST MONTH
Recurrently paying
users

Recurring
Payment % = 10%

2ND MONTH
Recurrently paying
users

Recurring
Payment % = 10%

3RD MONTH
Recurrently paying
users

ARPU = Monthly Fee = 100$

Revenue = 100$ * 100 = 10.000$

ARPU = Monthly Fee = 100$

Revenue = 100$ * 10 = 1.000$

ARPU = Monthly Fee = 100$

Revenue = 100$ * 1 = 100$

TOTAL REVENUE = 11.100$
TOTAL #USERS = 100

LTV = 111$

Chapter_06B

Measurement, analysis and testing base types (and associated concepts)

Head Metric

The Head Metric (by Löbchen & Fox) is nothing but the main metric of yours. The relevant literature uses many names for the same concept (e.g. One Metric That Matters, aka OMTM – by Croll and Yoskovitz; or Wildly Important Goal, aka WIG - McChesney, by Covey and Huling; etc.).

The literature agrees that this main metric has many essential features:

1. There is only one of it. To retain focus, you can only have one main metric.
2. Can be defined numerically. So its value can be precisely measured and defined.
3. It reflects your business goal. It's no accident that this is the main number. If the number shows a good value, you are successful. If not, you still have what to work on.

Note: The reason why I prefer the Head Metric expression out of these the most is its symbolism. Humans have one head which controls the entire body, but it still needs the rest of the organs and body parts to work well. The same hierarchy and cooperation can be seen between your business' Head Metric (the main metric) and the Body Metric (the subordinated metrics).
To reach your main goal all sub-goals – or at least most of them – need to be met (the same way all internal organs need to work for your head to work). Or if something's not right, you will immediately see it on the Head Metric (the same way you feel it in your head when you are sick).

Whichever expression you chose: always have a main metric! Otherwise, you will be watching too many analysis' and metrics and you will lose your way.

Note: I write in detail about the Head Metric in the Practical Data Handbook – although it has not been published yet… but you will know about it when it is.

Segment

A segment is a given part of your total target audience which you can separate based on one (or many) attribute(s). E.g. if you segment users based on gender, then you have a male and a female segment. If you chose location, it can be American users, European users, etc…
In the Chapters 2 and 4 we split users in groups from an activity and payment perspective. This was a kind of segmentation.

Segmentation

Splitting the audience according to certain attributes. This technique is very useful when used with other analysis'.
E.g. We want to measure the 3-Day-Retention % of our audience registering on the 1st of January. How many of those registering that day come back within 3 days. We can see that this ratio is 20%. Then we check this number segmenting mobile and desktop Users. And we see that 1% of mobile Users return, whilst 80% of desktop Users. We immediately know that something is not right with the mobile app (there's a bug or the product is just simply not practical to use on a mobile), but we're good on the desktop front. It's still a question though where to move from here (should we fix the mobile part or improve the desktop), but this is dependent on your strategy and a great CEO/PM/anyone will know the right answer.

A few typical segmentation types

- Based on the device (mobile/desktop/tablet)
- based on location
  - country
  - city
  - continent
  - etc.
- based on language
- based on gender
- based on age
- based on payment (explained in detail in CHAPTER 4)
- based on activity (explained in detail in CHAPTER 2)
- based on product preference
- based on the marketing channel
- based on the landing page
- etc, etc...

Cohort

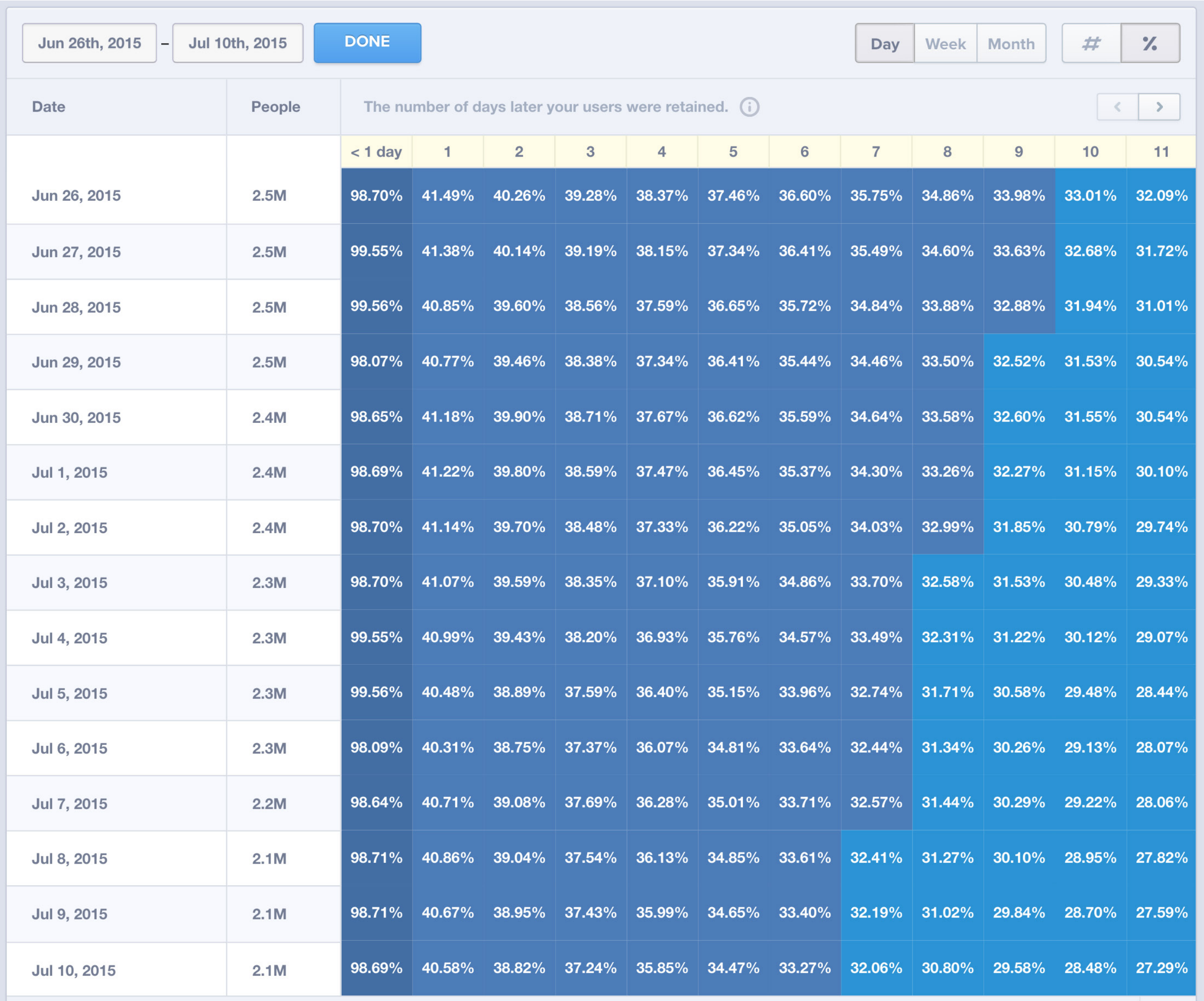
A cohort is ultimately a special segment type. A cohort is the splitting of users by time. So e.g. there is a cohort (group) for users who registered on 2016-01-01, a cohort for those users who registered on 2016-01-02, etc. But this can be the cohort of those making purchases in January, the cohort of those shopping in February... Anything. The main thing is to split the users into groups based on when they completed certain activities. In 99% of cases, this activity is actually the date of registration.

Note: Many people often incorrectly use the word cohort instead of segment. This doesn't generally cause any misunderstandings, but still…

Cohort-analysis

A special analysis type, the format of which roughly looks like this. This is a Mixpanel example, where the cohorts are split between the date of registration on a daily basis. (These are separate lines.) The date of registration is in the first column. The number of those who registered on a given day is in the second column. The rest of the columns show the percentage of return of the given cohort calculated within X number of days from registration – in other words, the ((Daily Active Users) / (Registered Users)) ratio within the given cohort – thus the X-Day-Retention Ratio.

If you want to learn more about this topic, I would recommend this article again:
http://data36.com/measuring-retention/



The number of days later your users were retained.

Date	People	< 1 day	1	2	3	4	5	6	7	8	9	10	11
Jun 26, 2015	2.5M	98.70%	41.49%	40.26%	39.28%	38.37%	37.46%	36.60%	35.75%	34.86%	33.98%	33.01%	32.09%
Jun 27, 2015	2.5M	99.55%	41.38%	40.14%	39.19%	38.15%	37.34%	36.41%	35.49%	34.60%	33.63%	32.68%	31.72%
Jun 28, 2015	2.5M	99.56%	40.85%	39.60%	38.56%	37.59%	36.65%	35.72%	34.84%	33.88%	32.88%	31.94%	31.01%
Jun 29, 2015	2.5M	98.07%	40.77%	39.46%	38.38%	37.34%	36.41%	35.44%	34.46%	33.50%	32.52%	31.53%	30.54%
Jun 30, 2015	2.4M	98.65%	41.18%	39.90%	38.71%	37.67%	36.62%	35.59%	34.64%	33.58%	32.60%	31.55%	30.54%
Jul 1, 2015	2.4M	98.69%	41.22%	39.80%	38.59%	37.47%	36.45%	35.37%	34.30%	33.26%	32.27%	31.15%	30.10%
Jul 2, 2015	2.4M	98.70%	41.14%	39.70%	38.48%	37.33%	36.22%	35.05%	34.03%	32.99%	31.85%	30.79%	29.74%
Jul 3, 2015	2.3M	98.70%	41.07%	39.59%	38.35%	37.10%	35.91%	34.86%	33.70%	32.58%	31.53%	30.48%	29.33%
Jul 4, 2015	2.3M	99.55%	40.99%	39.43%	38.20%	36.93%	35.76%	34.57%	33.49%	32.31%	31.22%	30.12%	29.07%
Jul 5, 2015	2.3M	99.56%	40.48%	38.89%	37.59%	36.40%	35.15%	33.96%	32.74%	31.71%	30.58%	29.48%	28.44%
Jul 6, 2015	2.3M	98.09%	40.31%	38.75%	37.37%	36.07%	34.81%	33.64%	32.44%	31.34%	30.26%	29.13%	28.07%
Jul 7, 2015	2.2M	98.64%	40.71%	39.08%	37.69%	36.28%	35.01%	33.71%	32.57%	31.44%	30.29%	29.22%	28.06%
Jul 8, 2015	2.1M	98.71%	40.86%	39.04%	37.54%	36.13%	34.85%	33.61%	32.41%	31.27%	30.10%	28.95%	27.82%
Jul 9, 2015	2.1M	98.71%	40.67%	38.95%	37.43%	35.99%	34.65%	33.40%	32.19%	31.02%	29.84%	28.70%	27.59%
Jul 10, 2015	2.1M	98.69%	40.58%	38.82%	37.24%	35.85%	34.47%	33.27%	32.06%	30.80%	29.58%	28.48%	27.29%

Funnel

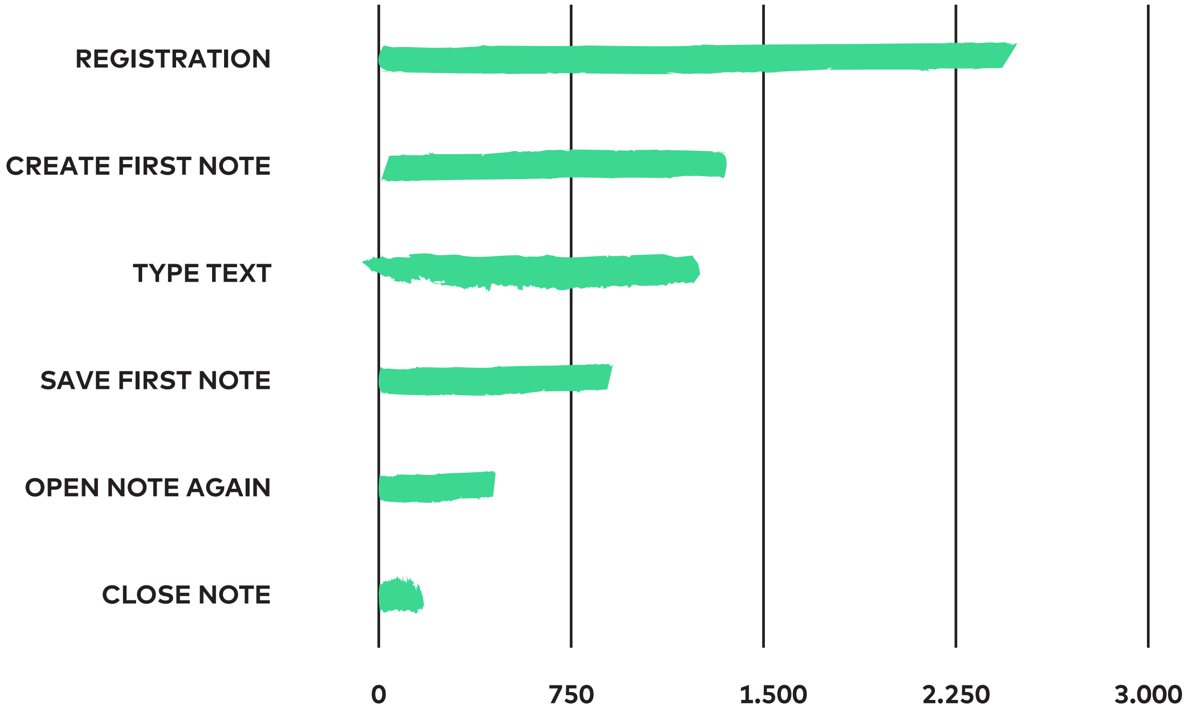
Generally, funnels are advised to describe strictly linear processes. The funnel itself is the path a User takes step by step from the beginning to the end of the process. The name comes from the shape of the related chart. During this process, more and more users drop out and fewer and fewer remain – visualizing this, we get a funnel-shaped diagram.

Funnel-analysis

Using this, we can examine the ratio of users dropping out at a step or advancing to the next one.

The easiest example is a registration form which most users fill in from top to bottom. It's expected that fewer and fewer users will fill in each field (the process is interrupted, like e.g. the boss comes in, the TV show comes on, the baby cries – or they just don't want to provide sensitive information like their bank details).

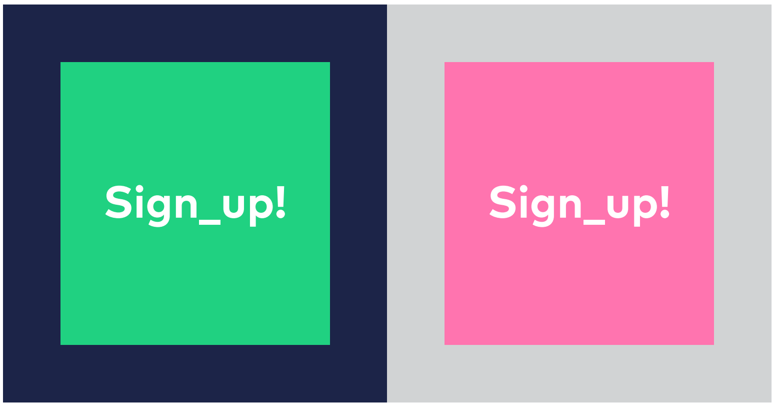
A well-visualized Funnel looks like this (e.g. in case of a note app):



If you want to learn more about this topic, I write in more detail about it in this article: http://data36.com/funnel-analysis/

AB-testing

The testing of two or more alternative versions of internet content. During the AB-test, when the User arrives to the page, they are automatically and randomly enrolled in a test or control group, so they see one version of our content. After this, we measure what they do on the page, and with what probability they reach the assigned goals.



With the right number of Users, we can use statistical methods to determine the most optimal version (usually the one that brings the most Revenue or activity.) A correctly implemented AB test has 5 important rules. These are:

1. Let the test and control groups (as similar sampling as possible) be determined at random!
2. Don’t allow the Users to know they are taking part in the test!
3. The different alternative versions should run at the same time!
4. Make the goal easily identifiable and measurable, so the results can be numerically defined!
5. Change one thing at a time!

A frequent question is what size sample should the AB-test be run on. This depends on many things. One is the baseline conversion of the control-version (e.g. Visit-to-Registration %= 3%). The higher this is, the smaller the sample should be. The other is the target performance growth (e.g. the Visit-to-Registration % should be 6%, that’s 100% growth). The higher this is, the smaller that sample should be. And finally, the targeted extent of the statistical significance (this is usually 95%, but for some it’s 99%). Based on the above, Optimizely created a great Sample Size Calculator which you can access

here:
http://bit.ly/opt-ssc

We talk in great detail about AB-Testing in our Data-driven Marketing Webinar:
http://www.data36.com/data-driven-marketing-webinar

Note1: The simplest and most often mentioned example for AB-testing is when on the page of an e-commerce shop the blue „Add to Cart" icon is coloured red (green, yellow, etc.) and they check how the different colours perform.

There are more complex AB-tests out there: Layout-tests, wording tests, title tests, creative tests on Facebook, etc... We provide numerous examples on this in our Data-driven Marketing Webinar.

Note2: Some sources split the so-called Multivariate-test from the AB-test. This is the playground of those with larger User-bases. The Multivariate-test works along the same lines as the AB-test. The only difference is that in the former we can change many things at the same time which can be combined with different variations with the different versions of the page. The results come out quicker, and the effects of certain elements can be discovered through various statistical methods.

Usibility testing

I'm not sure how this got into the data dictionary. Maybe because Usability testing as a qualitative research tool is a great and often necessary supplement to quantitative research.

Usability testing is damn simple. You invite a User into your office, you sit them by a computer and ask them to use your product. During this, you watch and take note of what they do. Ok, so it's not that easy, you need to keep in line with many rules in order to get relevant and useful information.

You need to know:

- Who is the test subject (it's worthwhile to pick this from your target group, if possible avoiding the designer and programming-orientated people)
- What's the scenario that needs to be completed (if any)
- What kind of questions you ask during the testing (so as not to influence the subject, it's good to ask open-ended questions)

and a few more little things a UX expert can tell you.

Note: As a Data company, we occasionally do Usability tests. This is for one reason: during these tests lots of problems, ideas and possibilities come up which we would never think of. So data analysis is simpler with few usability tests.

Chapter_06C

Additional valuable metrics

The above are the most often used metrics with the list of their relevant terminology. Of course there are many more types out there. One half of these are self-evident but special analysis'. For example:

- Cart Size
- Average Revenue per User
- Average Revenue per Paying User
- Average Revenue per Customer
- Click Through Rate
- Cost of Customer Acquisition
- etc.

If you don't happen to know these, I'm sure you can you can find a lots of information by searching on Google for a few seconds.

The other half are more difficult analytical methods. For example:

- Virality Score
- Score Carding
- Regression Analysis
- Clustering
- Principal Component Analysis
- Predictive analytical methods
- etc.

I didn't want to go too much into these in this minibooklet, as it would take up a crazy amount of pages, but I'm sure I'll get back to these somewhere else.

Chapter_07

Case studies

This chapter introduces in short how these concepts are used by companies in different business situations.

Chapter_07A

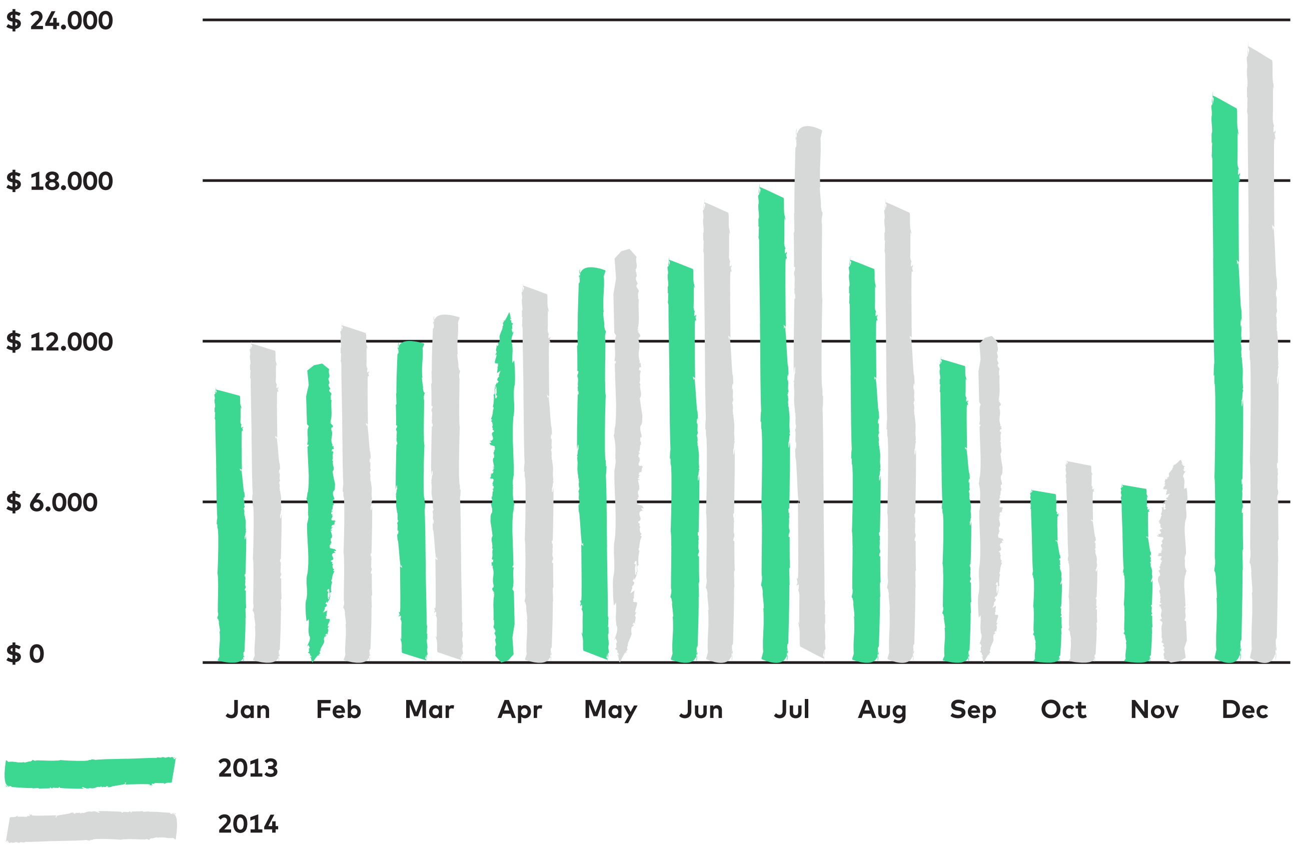
E-commerce case study - cohorts and segmentation

Note: unfortunately, the e-commerce sector is a really tough competitive market, so I could only write this case study by replacing the name of the company, product and numbers with something similar.

The Hiking Backpack E-Shop (if a so-called company does exist, apologies, I am not thinking of them, this is just a fictional example) began to analyze their data. They were curious about:

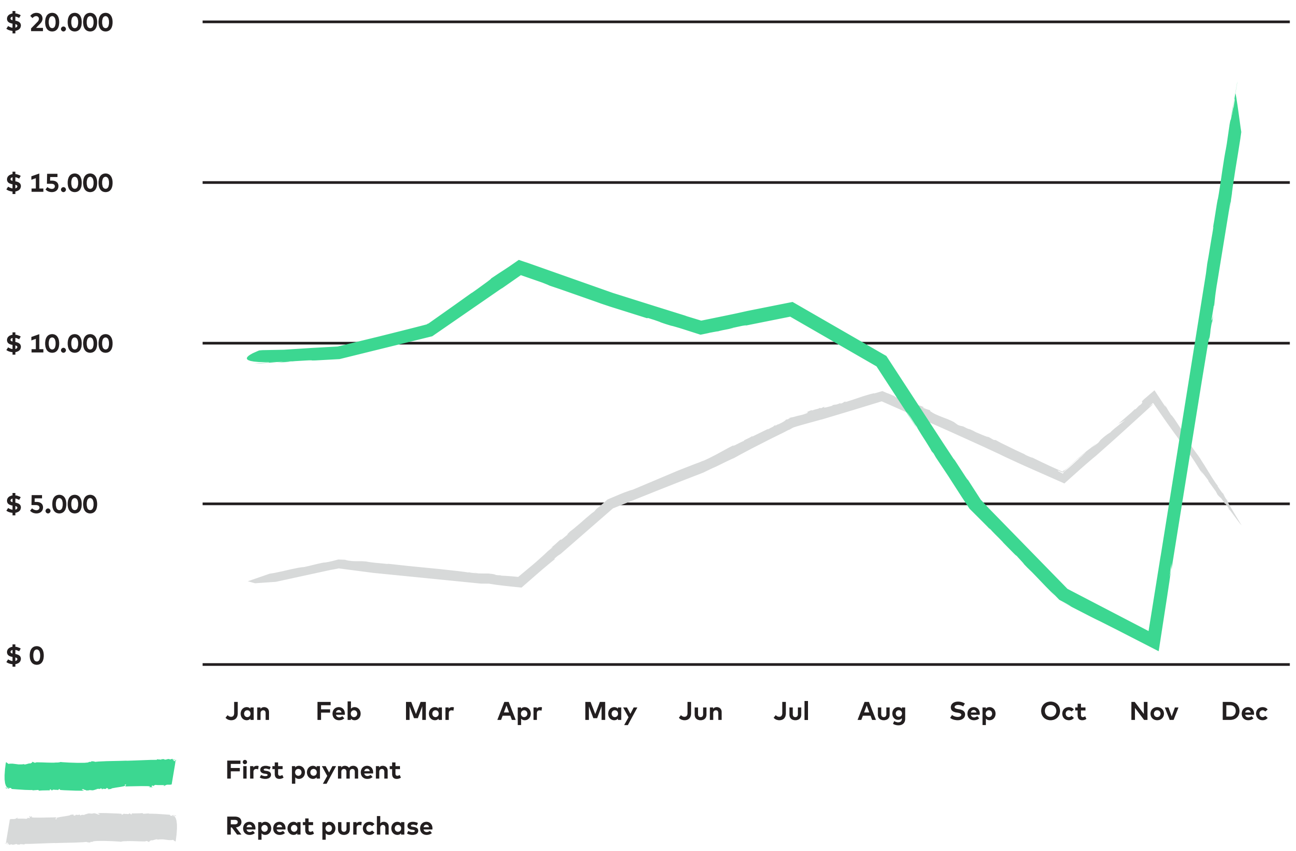
- Who is the best target group for them?
- What kind of product to offer to whom and when?
- Having answered these two questions, how can they reach the highest Revenue and higher Visitor-to-PremiumCustomer % in the long term?

The first thing they saw was that the sales performance fluctuates throughout the year.
This can be for a number of reasons of course, but knowing the circumstances we first thought that this is due to the nature of the product. To validate our suspicions, we looked at the 2013 vs. 2014 Revenue Chart on a monthly breakdown. The two years show a similar trend (we only see a small growth). We see the same for 2012 and 2011 as well.



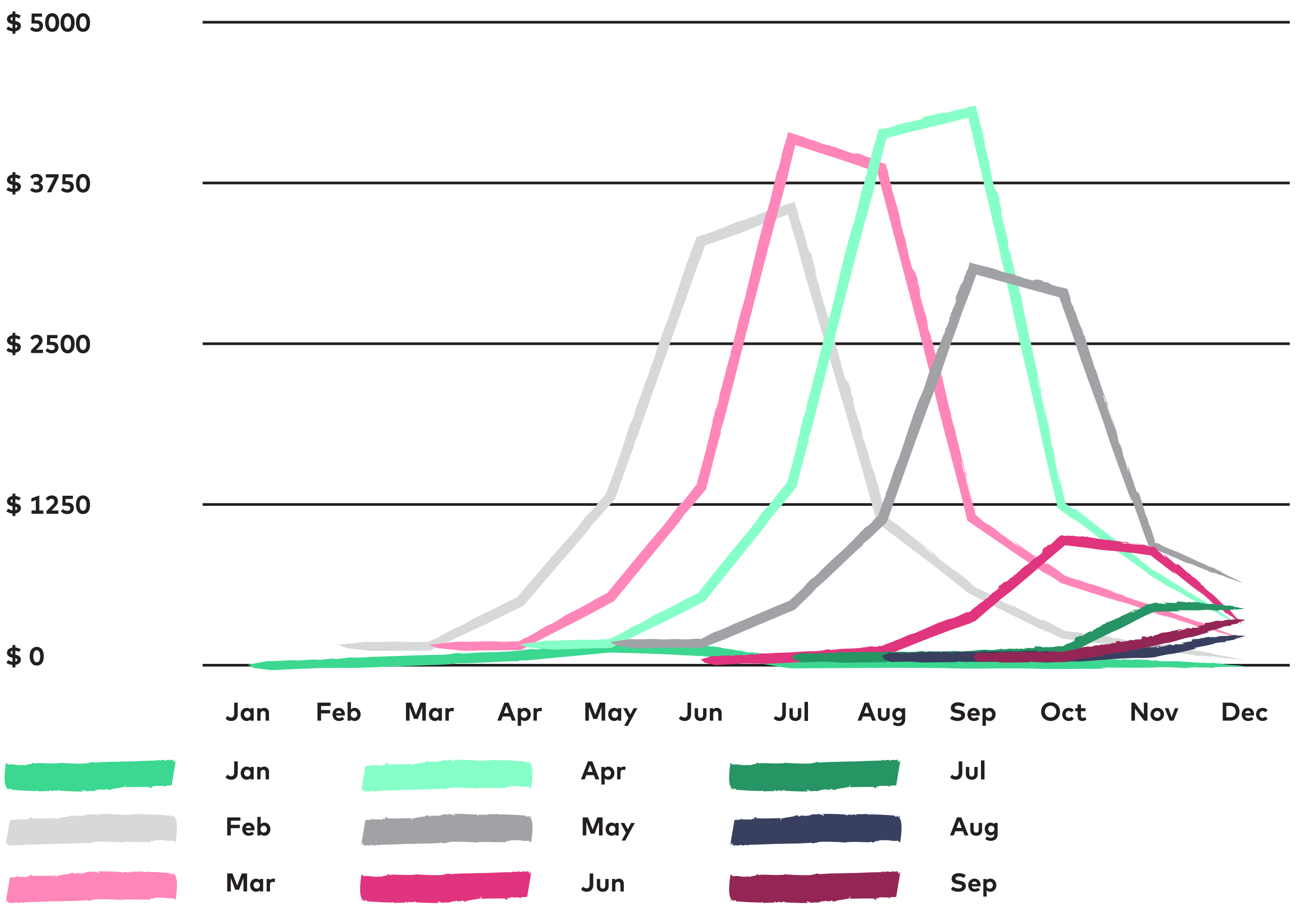
As can be expected, we did a number of User interviews and Usability tests, and checked some obvious analysis' based on different hypotheses. Most of these didn't give us any exciting results – but one of the segmentations had an interesting outcome.

We segmented the Revenue on the below chart based on Payment types. We can see that there was a constant change in 2014 on whether the „simple" First Payments (so namely the first purchase) or the Repeat Purchasse (when a previous Customer purchased again) brought in more Revenue.



It jumps out that the Revenue generated by New Customers drops in autumn, but returning Customers cover this gap.

In light of this, we created a Cohort analysis for those who made their first Payment in the shop in 2014. We looked at exactly how much was spent and when as a Repeat Purchase. We found this:



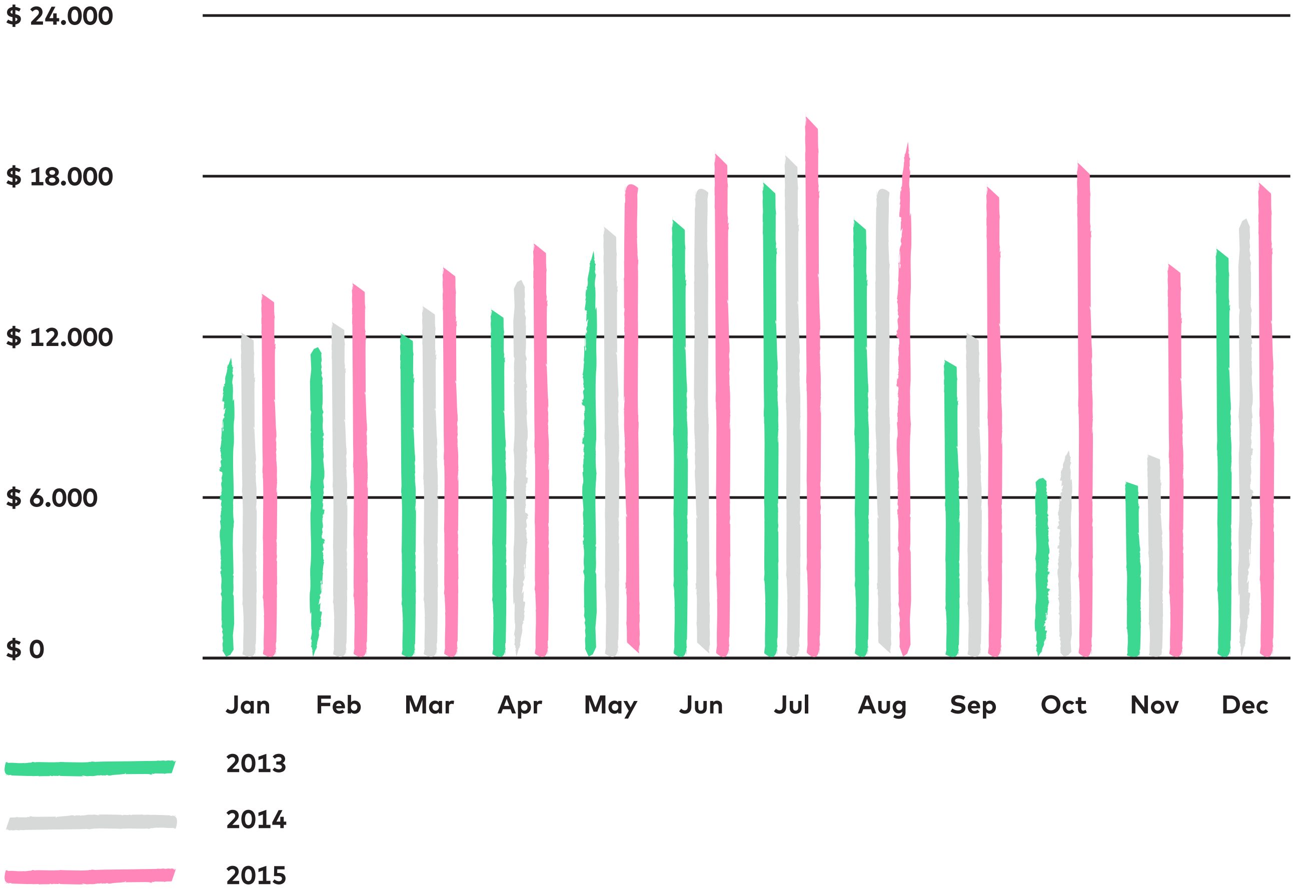
So the Customers from 2014 brought the best Revenue from a Repeat Purchase at the end of the summer and beginning of autumn. In fact, we also know that Customers from February, March, April and May are really strong and spend a lot 4-5 months after they make their first purchase (so July, August, September and October).

From this, two obvious reports followed.

One is to take a look at the same metrics, but through many years. (This also showed that the February-May Customers spend a lot as a Repeat Purchase. It is clear that it was them who took this seriously and planned their „trips" ahead and with that their „trip equipment". The rest shopped on an ad-hoc basis in the summer, or gave the backpack as a gift – typically around the Christmas period.)

The other is to define the exact product people purchase as a Repeat Purchase. This was a much simpler story. In short – they were able to find a well-targetable Customer Group and also what to sell them again and when.

The autumn campaign of 2015 was thus approached with a brand new strategy. Instead of aiming at new Customers, the current ones were targeted in these 3 months. This had its results.


$ 24.000
$ 18.000
$ 12.000
$ 6.000
$ 0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2013
2014
2015

Chapter_07B

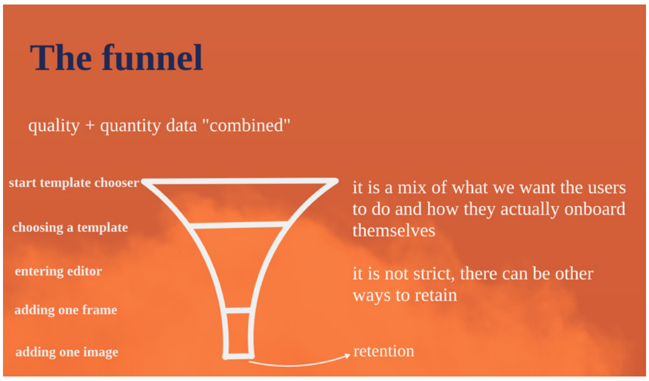
Funnel analysis at Prezi

Andris Balogh is the former senior Lead Data Analyst at Prezi. During his Prezi-years he gave an insightful presentation on how and for what he uses Funnel analysis.

Note: This was at the BData 2015 conference (organized by Data36).

„[...] When we have collected all the information from the analysis and have sat down with the Social Researcher and UX Researcher, we think over what kind of Funnels does a User have to go through to come back again (Retention). At Prezi, a Funnel is when a User goes into the Template Chooser where he/she picks from a Template, enters to the Editor and then starts to do other things. This is the structure which has come out of the analysis', Usability Tests and additional researches.

The Prezi Editor is not the kind of product in which you can only go down one path. Compare it to any other Registration process where you can't do things in another order, unless you give your name, email address, click on the registration button, click OK and then you get an email... In comparison, with the Prezi Editor a User can take many paths.



The funnel

quality + quantity data "combined"

it is a mix of what we want the users to do and how they actually onboard themselves

it is not strict, there can be other ways to retain

Due to this, a Funnel is a mix of what we want the User to do (based on which he/she understands the product), as well as what the Users actually do based on the analysis'.

So it's an interesting synthesis between expectation and reality. It's important to see that since this is not a strict Funnel, the User can come back in other ways, but we created a Funnel which mainly caps those who continuously stay Active Users. And those who at some point drop out will with most probability not return (Churn).

But what can you do with your Funnel? Definitely not starting to heal the top of the Funnel so more people can come in through there. It's not necessarily the best solution if you begin to fill the largest hole between two steps. I think the best option is if we begin our work at the bottom of the Funnel. Because if you begin to manically pack people to the top of the Funnel (e.g. with Google AdWords or Facebook Ads), those will drop out anyway. And those you load to the top and drop out will never come back. That's a wasted User.

So it's best to spend your time on those we know love us and have tried many of our products. Let's see what can help them and heal the bottom of the Funnel for them. You don't want to work with those who come and just take a peek at your product. So you gradually fix your Funnel upwards, and when it has reached a certain „thickness" where you say okay, this works, then you can start working on larger marketing costs and other good ideas. And bringing in the Users.

Let's look specifically at the case of Prezi. In this case, placing the first image in the Funnel was the most important part. This is a real decision: the development of things begins with image placement!

This means that the Developer, the UX Researcher and the Designer sit down and begin to work around

this function. During this, there is ongoing analyses of course, as it’s better to pin-point what is the exact problem. Usability Tests can change into something that only deals with image placement. Also, the analysts can create a higher resolution for the part of the Funnel where there is an issue.
Simply speaking, we place a sub-funnel into the place where images are placed. E.g.

1. They press the „add image” button, then
2. They click on „choose image from computer”, then
3. The image is uploaded to the server, then
4. It’s uploaded to Prezi.

And this way, we can easily see where the problem is.”

Chapter_07C

AB-Testing at Ustream

Gergely Schmidt is a Product Manager at Ustream. He also presented at the BData conference on how AB-testing works at their company. Here's a short extract:

„One of our products is about how you can purchase Ustream Pro Broadcasting and what kind of extra features you will have. One of the most important is that you can broadcast to your viewers advertisement-free. This entails a registration form where we ask about pretty much everything about you. We tried to optimize this page so we can have as many subscribers (supplement: Recurrently Paying User) as possible.
The test was about whether to have an overview page where Users can check what kind of data they have provided (A-version) or not to have such a page (B-version). At the bottom of the form (in both versions) there was a Complete Purchase button as well where we showed the Users how much they will have to pay. Interestingly enough, there was not a big difference in the number of purchases. We stood there surprised, thinking we did something wrong. But we didn't. But we noticed much later – which was not even measured in the original testing – that the number of Refunds differed. Those who received the overview form requested a Refund much less than those who received the shortened form, as those asked for their money back more often. So we only realized way after the testing that from this perspective, the overview form version was the winner. From this, you can gather that it's important to follow up on every test which you run on your site, as it may not be influencing the metrics you initially worked on."

Chapter_07D

Usability testing at Skyscanner

Laci Kardos, one of the Product Managers at Skyscanner explained in a Data36 interview (http://data36.com/product-research-interview-product-development-at-skyscanner/) how „codeless testing?" works and why it's good. Here is – in my opinion – the most useful part of the conversation.

Tomi: *"How should we imagine codeless testing?"*
Laci: *"Just imagine a simple wireframe-featured prototype. We create screens and we link these together. It's very important for the rhythm of the tests to provide a base rhythm to the entire product development. If we meet a user, we want to show them something. We give them a prototype, and the researcher's job is to do the test. It's in the basic interest of the team to be at as many testings each week as possible. Since it's not just important for the designer, the product manager or the researcher to see whether what they have created works, whether it's valuable, usable, but it's also crucial for the developer, too. These are generally 30 minute tests. Sometimes they are built upon scenarios. For example, „Imagine that you want to travel and you start to use the app you have downloaded" – on iOS, Android, a tablet or on a mobile. During the user test we can see where the process halts – during this we speak to the tester to understand the „why's". Then we speak to the team and go through what we have learned, what we heard. As before this, we had certain presumptions, and following the test these are either verified or not. It's at times like these when we see what doesn't work, what works really well and sometimes we even see things we did not expect. In my experience the value and utility of a product can be judged after 3-4 tests."*

Conclusion

Thank you for taking the time and energy to read this booklet. I know it's not a simple topic and – unless someone is a data-fan like some of us – it may be a dry read at times. But I tried to write it in an interesting way.

I hope you can make use of what you read in practice and create a consistent and thought-out common language on data in your organization. As I mentioned in the introduction chapter, the goal is not to have an 100% match with what is written here, but more to give you some inspiration and ideas!

I wish you good luck and great success!

Contact

If you have any questions with regard to this booklet – whether you found a mistake, a typo or you had a great idea (or you would do something differently) – write to me to this email address

tomi@data36.com

Also don't forget to follow me on Twitter

https://twitter.com/data36_com

Or subscribe to the data36 newsletter, if you did not so far

http://data36.com/datadictionary

Note: A big thanks to those who reviewed, gave their thoughts on and supplemented the booklet before the first edition! Especially to Andris Balogh, Agoston David, Gabor Papp, Adrian Sandorfy, David Szabo and Attila Virag!

Graphic design by Faraway.hu