Data Dictionary

Written_for_Online Businesses

Tomi_Mester

Created September 2015

Why_is_this important?

Created by
Tomi Mester

Follow me on Twitter:
@data36_com

When an online business starts to use data, they usually read a bunch of articles and books on the subject. In the best-case scenario, they hire 1-3 data scientists, set up a data infrastructure, and come up with a data strategy. Slowly everyone at the company starts to use data and an awesome data-driven organization is born. Hooray!

But along the way there will be some confusion. Data science is not a set-in-stone kind of science (yet). Various books and articles will give you various strategies and various naming conventions. It's not uncommon for the same concept to have 3-4 different names in different places. Even worse is the other way around: sometimes, the same word is used for different data concepts.

The more clients I worked with at Data36, the more problematic this issue became. So I decided to create a dictionary which unifies the most important data expressions and places them within a clear framework. The main points were:

- consistency
- simplicity
- clear and easy-to-use naming conventions

This is how Data36's Practical Data Dictionary came about. I originally gave it to my clients only. But I decided to open-source it - because I've seen that these issues concern almost every online business.

Note 1: This summary is general. Some processes in your product/service may not follow one another in the same sequence as I describe in this book. Some parts might be irrelevant to you. No worries – this is not a rule book, it's just a guide. I trust you there. You are smart: pick and choose the parts that are useful for you!

Note 2: Chapter seven consists of case studies. As you're going through the dictionary for the first time, you can refer to it to understand how these concepts are applied at real companies.

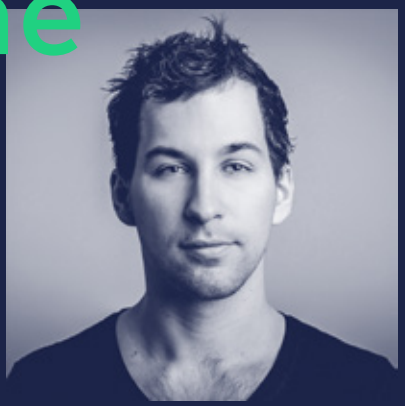
Content

Hi, I'm Tomi Mester, a practicing data analyst and researcher for 7+ years. I've worked for Prezi, iZettle and several smaller companies as an analyst or consultant.

I'm the author of the Data36 blog where I write posts and tutorials on a weekly basis about data science, A/B testing, online research and coding. I'm an O'Reilly author and presenter at TEDxYouth, Barcelona E-commerce Summit and Stockholm Analytics Day.

Find more info here:
My LinkedIn profile: https://www.linkedin.com/in/tomimester
My Email address: tomimester@data36.com
Follow me on Twitter: https://twitter.com/data36_com

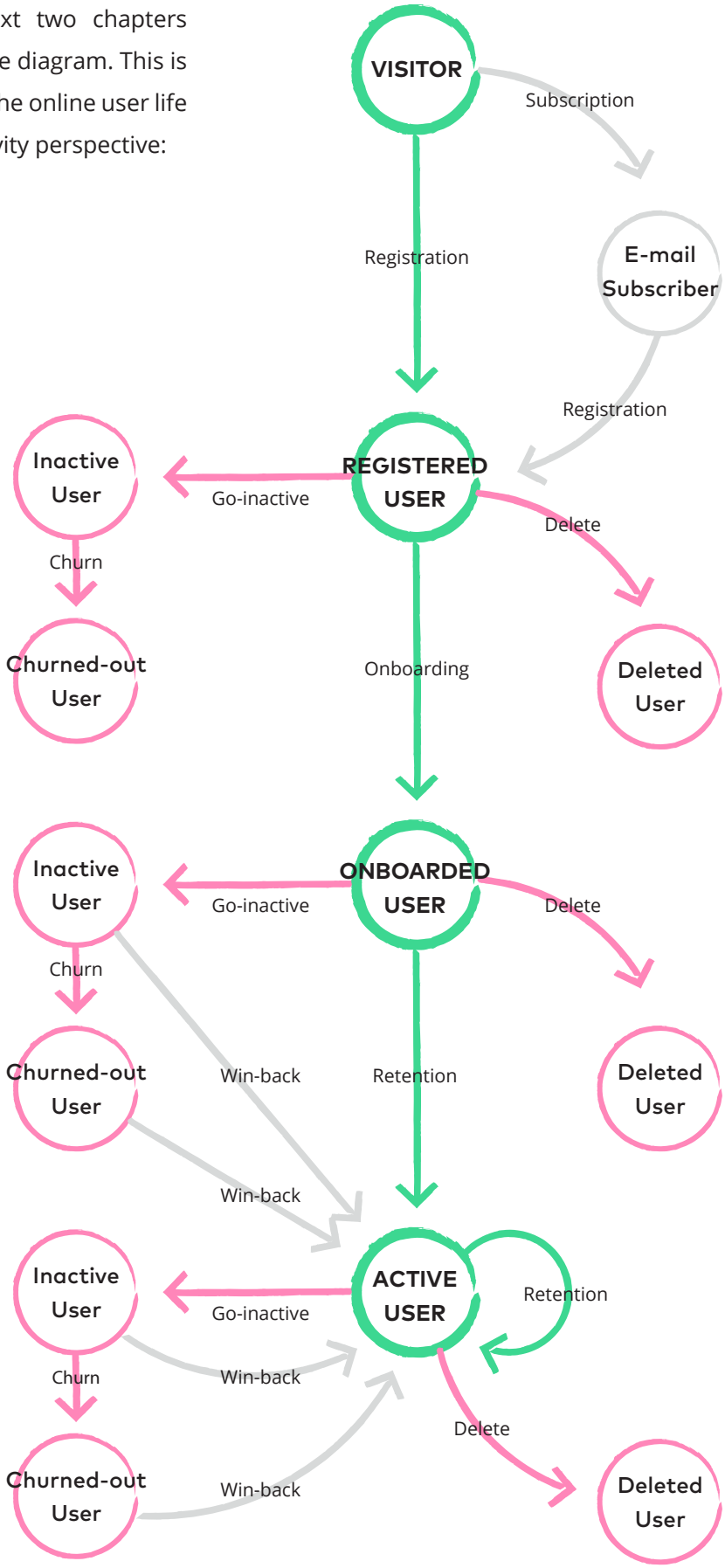
About me



Chapter_01

User activity - The main events of the user life cycle

Here are the next two chapters summarized in one diagram. This is a general view of the online user life cycle from an activity perspective:



Visit	When someone visits your website.
E-mail subscription	When someone visits your website and provides his/her email address – but may not necessarily register for your product/service. (Most commonly: signing up for the newsletter.)
Registration	When someone visits your website and registers, so he/she can come back and login later. (Most commonly: providing an email address and setting up a password.)
Onboarding	This usually (preferably) happens right after Registration. During Onboarding, the Registered User goes through the key steps of your product. During the Onboarding the User becomes familiar with the main value that your product offers *(e.g. for an invoicing software, he/she creates and sends the first invoice, etc...)* If you haven't defined your Onboarding process yet, do it as soon as you can! When you do so, keep in mind: when a user is onboarded, she has to clearly see the value of your product/service, to increase the chances that she'll start using it regularly (e.g. do her monthly invoices with it).
Retention	Retaining users means keeping them active. An *Active User* will use your product/service again and again! Note: If the user logged into her user account, it does not necessarily mean that she used your product. You'd be surprised to see the logged-in-but-did-nothing-else user ratio on many products... So when you measure activity tie it to the end of your Onboarding process. (E.g. a user logged in to your invoicing software doesn't count as an Active User. A user who sent an invoice does.)
Go-Inactive	When a user does not use your product/service for a given time period (say, a week).
Churn	When an Inactive User does not use your product/service for a given time period (say, four weeks -- so it would take five weeks for a user to go from Active to Churned).
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 activity - Naming conventions for the different user segments

Visitor	Someone who visits your website, a potential Registered User.
E-mail Subscriber	A visitor who provides their email address (but doesn't register).
Registered User, (more commonly: User)	The kind of Visitor who registers, that is, provides their email address (or any kind of unique identifier) for which you create a user account.
Onboarded User	A User who has gone through your Onboarding process and (preferably) understood your product/service.
Active User	A User who has used your product within a specific time-frame that you have defined (e.g. a given month, week or even day). This is a status of a user that will change all the time. Note: Again! If the user logged into her user account, it does not necessarily mean that she used your product. You'd be surprised to see the ratio of the logged-in-but-did-nothing-else user ratio on many products... So when you measure activity tie it to the end of your Onboarding process. (E.g. a user logged in to your invoicing software doesn't count as an Active User. A user who sent an invoice does.)
Inactive User	A User who has not used your product for a specific time-frame that you have defined (e.g. a given month, given week, given day or given hour). This status can change all the time.
Churned-out User	A User who has not used your product for a specified, lengthier time-frame that you have defined (e.g. two months inactivity time). This status can change all the time.
Deleted User	A User we deleted from our system or who has deleted themselves.

Note1: Check the diagram above again! You will see that these user types are in fact different statuses of your users. The Email Subscriber, Registered User and Onboarded User statuses are one-time statuses. The main goal is to "push" your Users through these – as quickly as possible. Then to keep them as Active Users for as long as possible. This will not work with everyone of course. But in that sense, your goal is to have a relatively low number of Email Subscribers, Registered Users and Onboarded Users. (Early user stages.) And you want most of the Users to be Active users. The rest will be coming and going between the Active/Inactive/Churned-out status.

Regardless, it's important to have the E-mail Subscriber/Registered/Onboarded segments in your database since these Users are new to your product - and curious. So there is a higher potential in working with them and focusing on them.

Extra for Chapter_02

User activity - Subsegments

When running in-depth data research projects, you aren't only interested what phase your users are in right now (Onboarded, Active, etc.). You need to know what phase they were in before. It makes a difference whether a currently Inactive User only registered before going inactive (and hadn't tried the product yet), was an Onboarded User (tried the product, but only once), or had been an Active User. Let's segment your users from this perspective as well!

INACTIVE USER SEGMENTS

Registered-then-Inactive User

A User who after Registration immediately became an Inactive User.

Comment: Another common data term is “Dead-On-Arrival.” But I don’t like it.

Onboarding-then-Inactive User

A User who, after Onboarding, immediately became an Inactive User.

Active-then-Inactive User

A User who was Active, but then became Inactive.

ACTIVE USER SEGMENTS

Onboarded-then-Active User

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

Active-then-Active User

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

Inactive-then-Active User

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

Churned-then-Active User

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

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

Note2: At the same time, if you create too many subcategories, you can lose focus by focusing on too many segments.

Note3: Since we touched on the topic of focus... It’s an important strategic question to decide on which of the above categories (8 + 3 + 3 + your own subcategories = 14+) you will focus on. A lot of articles explain why it’s better to pay attention to Registered Users rather than 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. So instead of reading articles, run analyses and decide based on that what’s important for you. And put that segment at the center of your attention.

Another
Supplement to Chapter_02

User segments by timeframes

Segmenting Users by activity is more actionable if you break them further down by timeframes (e.g. Daily Active Users). Based on my practical experience, it's better to define these as absolute timeframes rather than relative ones.

This means that when it comes to analysis, your Daily Active Users are not the Users who were active *in the past 24 hours* (that would be a constantly changing group!), but those who, for instance, were active between 2016-01-01- 00:00 and 24:00 (this is a fixed group; as soon as 2016-01-01 24:00 has passed, the group of users in this segment does not change).

You have to define your own segments, and you can combine these with the naming convention we discussed above. A few examples:

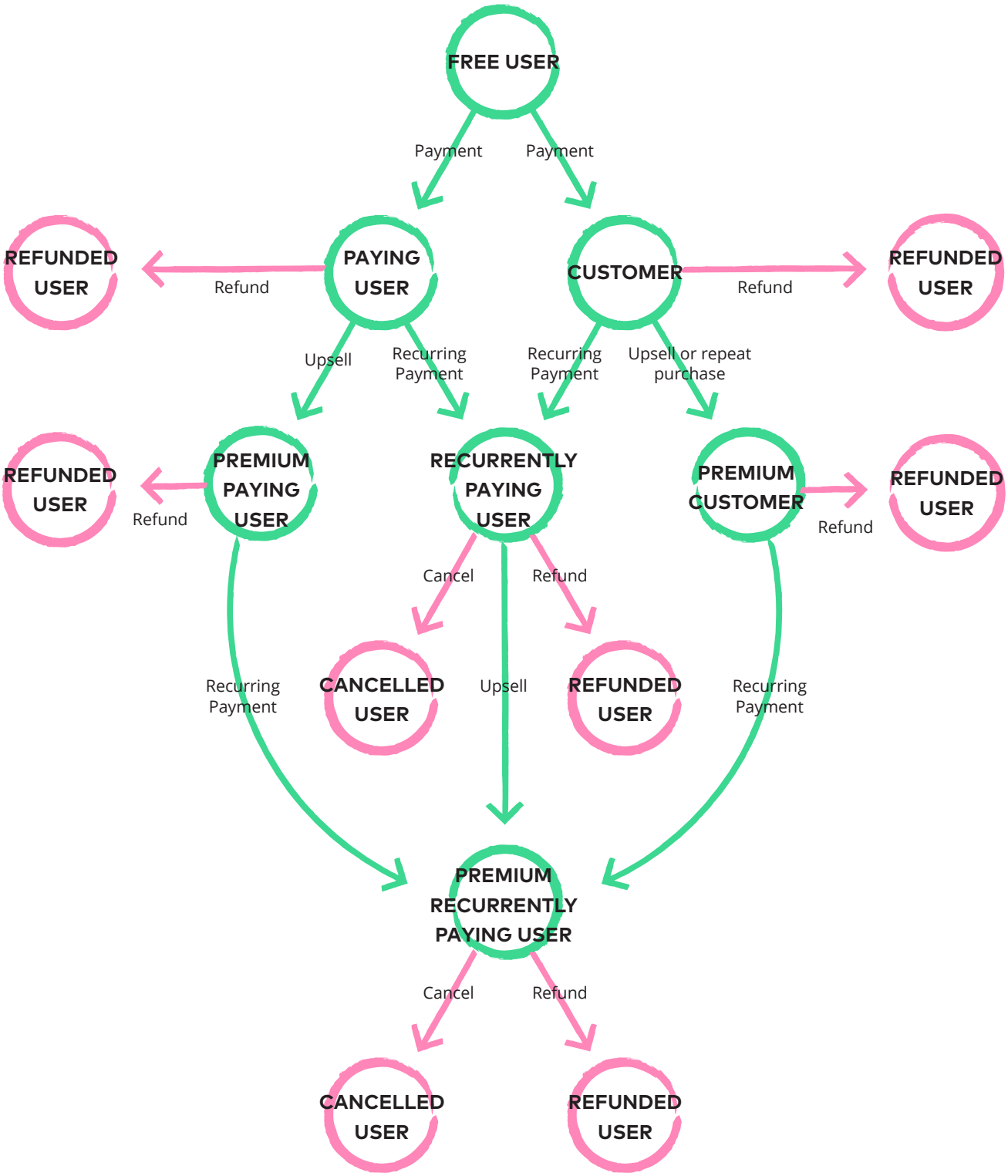
- Daily Active Users (e.g. the number of Active Users on 2016-01-01 is: 352)
- Weekly Onboarded Users (e.g. the number of Onboarded Users on W1 of 2016 is: 1.860)
- Yearly Churned-Out Users (e.g. Churned-out users in 2015 is: 21.512)
- and so on..

Chapter_03

Payments - the main events of the user life cycle

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

Note1: Payment models can be highly varied. This is the general picture, but for your business, it's likely that only a small part of it will be relevant.



Payment	A transaction. When someone pays you. 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 - after receiving their money back - are usually a very satisfied segment.
Recurring Payment	Payment that automatically renews and charges your user monthly or yearly. Most common with services, but it can apply to products, too (e.g. a magazine subscription).
Cancel	Cancellation of the Recurring Payment. 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 - but it's different at its core. Selling a Customer one more thing. (A new pair of shoes, or another pair of what he's already bought.)

Note2: Keeping a consequent and straightforward naming convention is extremely important for payment types. As you can see, there are quite a few similar-looking payment-related events; but in practice (how they impact your business model and your revenue) they are very different.

Chapter_04

Payments - the naming convention for the different user segments

Free User	A User who has registered, may be using your product or service but has not yet made payment to you.
Customer	Someone who has purchased at least one product from us. Not the same as a Paying User!
Paying User	A User who has paid to use your service for a given time period (e.g. to unlock a premium feature of your product). Not the same as a Customer! Note: The main difference between a Paying User and a Customer is this: a Paying User pays for a service for a given time period (this can be renewed - actually, it's usually a Recurring Payment), whilst a Customer pays for a specific product once and can use it forever. E.g. if someone buys Microsoft Office 2015 in a one-time payment, then she is a Customer, but if she subscribes to Microsoft 365 and pays a monthly fee for using it, she's a Paying User.
Refunded User	A User or Customer who for some reason asked for her money back (and received it). (E.g. the Customer did not like the purchased shoes and sent them back; or the User did not like the software she subscribed for, cancelled and asked for a refund.)
Cancelled User	A User who was a Recurrently Paying User, but eventually cancelled her subscription. (But did not necessarily ask for a refund).

Extra

Other payment-related user segments

Note 1: Don't forget to go back to the diagram on the previous page -- so it'll be easier to understand visually where these segments fit!

Premium Customer	A special Customer segment that spends significantly more than average Customers. (Note: you should define an exact value!)
Premium Paying User	A special Paying User who spends significantly more than average Paying Users.
Recurrently Paying User	A Paying User who subscribed for a service (or sometimes for a product) and whose payment is automatically renewed every month or every year - until Cancellation.
Premium Recurrently Paying User	A Recurrently Paying User who spends significantly more than the average Recurrently Paying User.

Note 2: Naturally, these groups can also be further divided. Premium Customers, especially, can be split into different levels. Also, Recurrently Paying Users can be graded based on the number of renewed payment cycles. Don't create too many small segments though. You want to avoid confusion.

Note 3: Above which value does someone become a Premium category User? This needs to be defined by you. Rule of thumb: take the top X% (e.g. top 10%) of your Paying Users.

Chapter_05

Summarizing everything so far

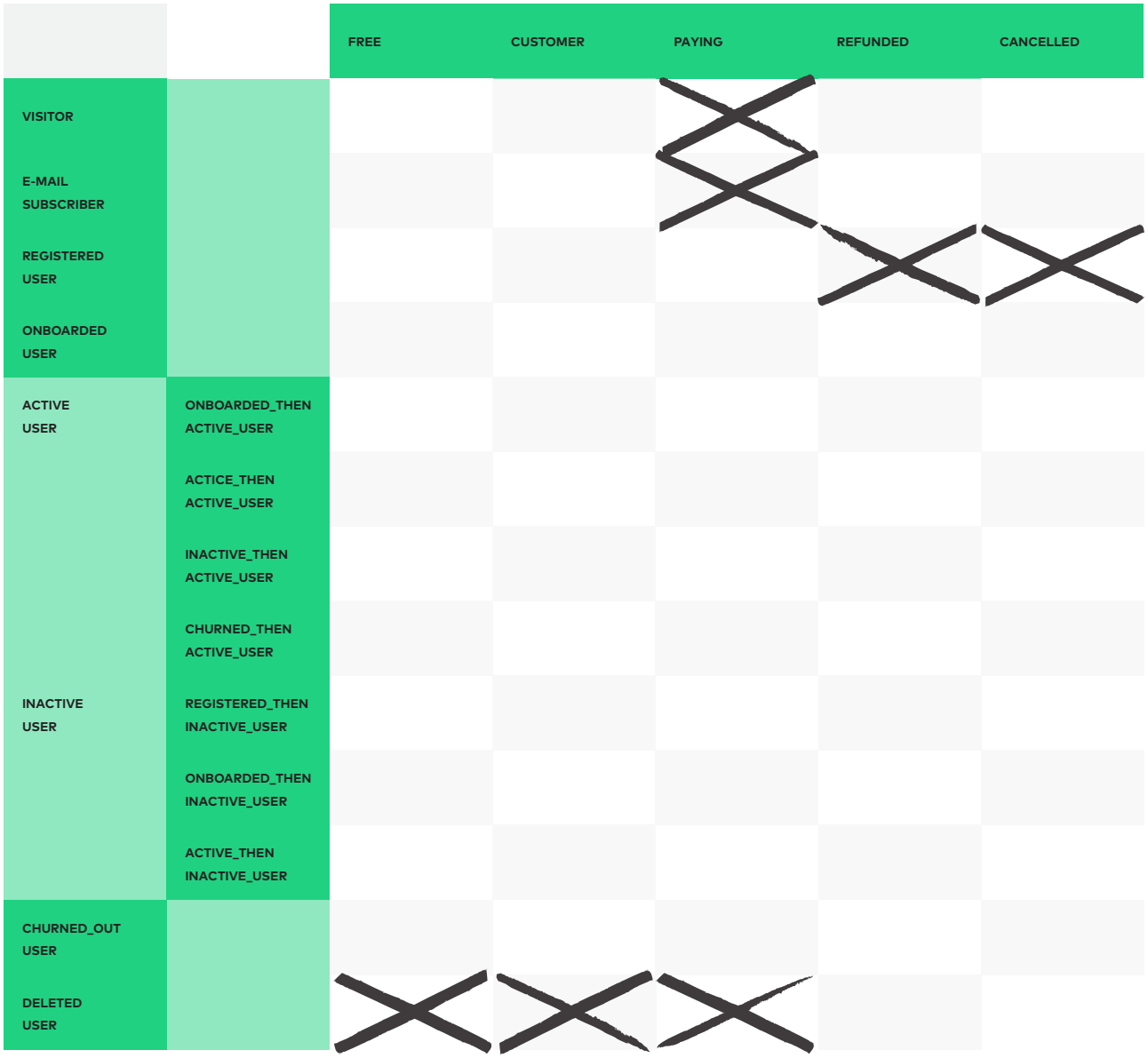
I collected into one diagram all the different User segments based on activity and payment.
There are many categories in it. If you remove the impossible categories (e.g. the paying user who is not registered), we still have 58 groups.
It can be further expanded with your own categories. At big online businesses (500+ employees), it's possible for each group to have its own marketing and product strategy... But if you run a smaller business, then it's very important to find your focus.

Consider this when you choose your focus segments:

- Which segment has the most users/customers?
- Which segment is the most problematic?
- Which segment has the largest potential?

Note: Read this article to learn more about Data Science for Business:

https://data36.com/data-science-for-business/



Chapter_06

Metrics and analytics method

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

Important ratios

"X"-Day-Retention

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

Note: The value of "X" has key importance, yet it is a very difficult value to define. Four principles can help you with the definition. The first principle is the "own-expectations" principle: you define how often you expect users to return based on your service's features. (E.g. for a social media app you can expect daily retention, but for a flight-search service, healthy retention can be up to 6 months.) The second principle is the data-centric principle: check the frequency of return based on your existing users' data. The third is the "quicker-the-better" principle: it's easier to measure and it's a better target if your users come back as often as possible. For this reason, if you are unsure of whether to make the target retention time-frame 3 or 4 days: always pick 3. The fourth is the others-know-already principle: look for benchmarks in your own market. I dive deeper into this topic here: http://data36.com/measuring-retention/

Retention %

The ((Active User)/(Registered User)) rate. (As you know, an Active User is someone who has used your product more than once within the X-Day Retention time-frame.)

Leave %

The ((Inactive User)/(Registered User)) rate. (Similarly to the previous point: An Inactive User is someone has not used your product within the X-Day Retention time-frame.)

"Y"-Day-Churn

The time-frame within which an Inactive User needs to return to not become a Churned-out User. The "Y-Day-Churn" value is usually not too far from the "X-Day-Retention" value. (e.g. if your retention timeframe is one week, then your churn timeframe is probably one month or so).

Churn %

The ((Churned-out User)/(Registered User)) ratio.

Win-back %

The rate of users who went Active after being Inactive, compared to the number of those who don'T. (Usually we calculate this ratio for specific Win-back campaigns.)

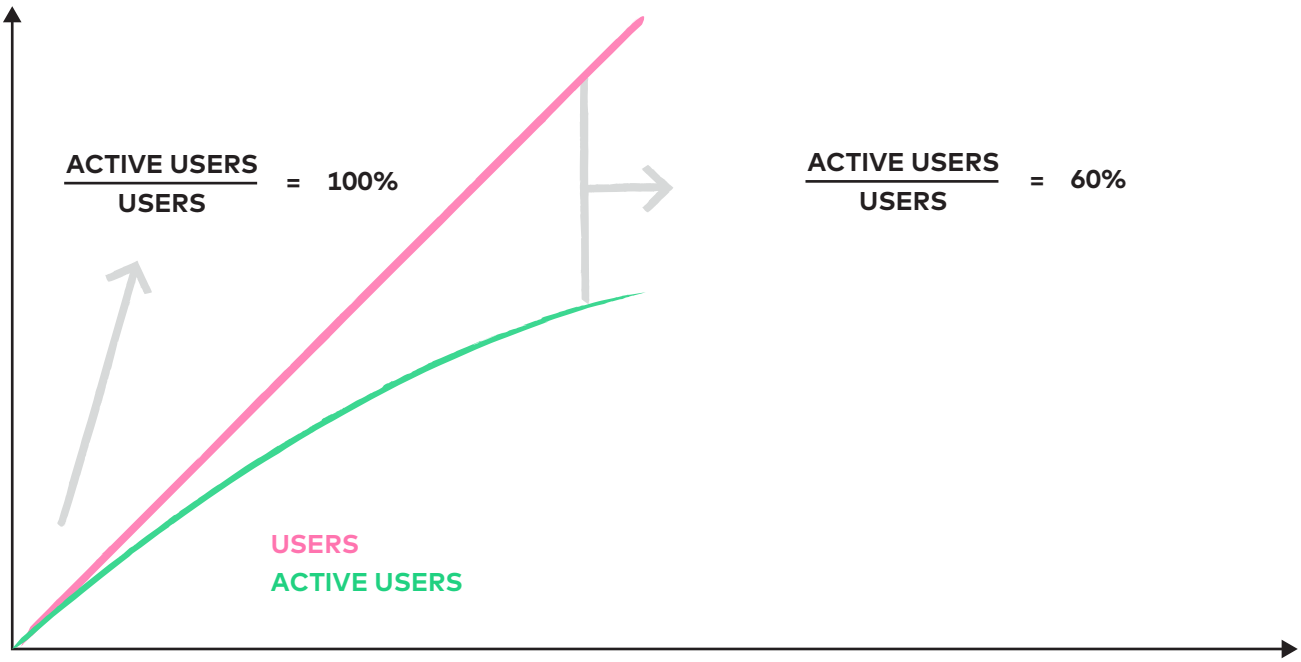
Registration %

The ratio of ((Registered Users)/(Visitors)) on a given day (or week or month).

X-to-Y %

Following the above examples, any ratio between two user activity statuses can be calculated.

Comment: Be smart with choosing your timeframes! Again! Use cohorts (see details below); otherwise you can easily mislead yourself. And be sure what you're measuring makes sense. Let's take a simple example: the (Daily Active Users) / (All Users) ratio will inevitably decrease over time. During the first few days of the product launch, most of your Users will be Active Users. Later, as more and more Users churn out, this ratio will constantly decrease. This is normal, but it also means that the (Daily Active Users) / (All Users) ratio will not be informative at all.



Conversion %

Although this is a common expression, we don't use it often with complex products. It's too general. "Conversion" can mean the performance of an advertisement, a purchase, a registration. Anything. It's difficult to use it in a meaningful way within a company. Or in a data dictionary!

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, because it's usually highly correlated to profit but it's a good deal easier to measure.

Note 1: In more complex analyses, we do calculate profit. It's as simple as deducting the costs from the Revenue. The real difficulty is in estimating the real cost of a project - e.g. considering salaries, alternative costs, travel expenses, etc... Usually it isn't worth it to bring all these into your analyses.

Note 2: Revenue is not just calculated on a company level, it can be done for subsegments or per product, too! See "Segmentation" and "Case Studies" below.

Repeat Purchase %

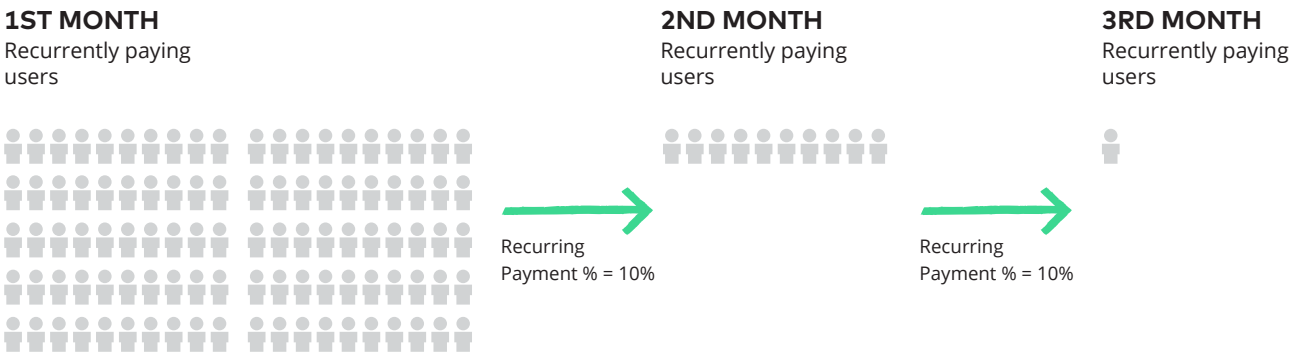
The probability of a repeat purchase from a customer (provided you have something to upsell or sell again).

Note: For simplicity, I usually put cross-selling (when you sell a product with another product) into this category as well. (e.g. movie tickets and Coke.)

Recurring Payment

(Similarly to the % of a Repeat Purchase) it gives the probability of a Paying User to keep paying for your service. (Remember, Recurring Payments are often automatically charged.)

Here's a simple example. Let's say you have a service with monthly, automatically renewed subscriptions. On average 90% of your users Cancel their subscriptions. It means that the Recurring Payment % is 10%. In other words: out of 100 users, 10 will pay for the second month again, and out of those 10 users, 1 will pay for the third month. (This is simplified, of course).



Lifetime value (LTV)

LTV gives the average revenue generated by one User during their entire lifecycle (the whole time they are an active and paying user). This value is incredibly useful for the calculation of profitability – and within that, the calculation of your maximum costs. A simple advantage of knowing your LTV numbers: it makes it easy to calculate if it's worth spending "X" on a given paid ad campaign which brings "Y" number of "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 before you can see expected profit.

The problem is that LTV is not easy to calculate precisely. Sometimes, even a Churned-out User can come back after 2 years through some miracle – and can start generating Revenue out of nowhere... How can you account for that your Lifetime Value calculation?

The right LTV calculation method depends on your business model. You can find many tutorials on how to "calculate lifetime value" on the Internet. Read these carefully but with a critical eye towards whether they are the right fit for your business.

Once you have found a suitable LTV calculation method, verify if the results are realistic. If yes, you're good.

Anyway:

I'll show you a simple and relatively good model here, which uses only two values: the Average Revenue per User and the Repeat Purchase %. Here's the formula:

Lifetime Value = Average Revenue per User * (1 + (Repeat Purchase%) + (Repeat Purchase%)^2 + (Repeat Purchase%)^3 + (Repeat Purchase%)^4 + (Repeat Purchase%)^5 + (Repeat Purchase%)^6...)

For example:

Average Revenue per User = $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. It's better to be pleasantly surprised rather than disappointed, right?

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

Most common analytics methods and important metrics

Your Most Important Metric

Different books and articles use 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.).

They agree that this main metric has many essential features:

1. To keep your focus, you can only have one most important metric.
2. It has to be measurable - and clearly defined.
3. It reflects your business goals and your users' success.

Whichever expression you chose: always start your data project by defining your most important metric! Otherwise, you will get confused and eventually lose your way.
Recommended article: https://data36.com/important-metric/

Segment

A segment is a specific part of your total audience that you can separate out based on one or more attributes. E.g. if you segment users based on device types, then you have a desktop, a mobile and a tablet segment. If you chose location, it can be users from USA, users from Europe, etc...
In Chapters 2 and 4 we split users into groups from an activity and payment perspective. That was a kind of segmentation, too.

Segmentation

An important analytics method. Splitting the audience into segments. Here's a simple example of how it can be useful for you:
You run a quick analysis and calculate the 3-Day-Retention % of your users who registered on the 1st of January. The question is: How many of them come back within 3 days and use your service/product again? Let's say you find that this ratio is 20%. Then you segment the registered users by device type. Now it gets exciting: you see that retention is 1% for mobile users and 80% for desktop users. Boom! You immediately know that something is not right with the mobile app (there's a bug, or the product is simply not designed for mobile), but you're doing great on desktop. It's still an open question, though, what's the next step from here: should you fix the mobile issue or start

focusing on desktop...? That depends on context and your strategy, but at least you have the data at hand.
Recommended article: https://data36.com/data-beats-opinion/

A few typical segmentation types

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

Cohort

A cohort is sort of a special segment type.
The essence of cohort analysis is that it groups users who share common characteristics (or experiences) within a defined time-span. Easiest example: a cohort can be all users who registered for your service on the same day (e.g. 2015-02-07).
But a cohort can be created based on making purchases, visiting the website, clicking a button... Anything. The main thing is to split the users into groups based on when they completed certain activities.

Note: People often use the word 'cohort' instead of 'segment'. That's just incorrect!

Cohort analysis

A special analysis type - best for measuring user retention. See a Mixpanel example here! The cohorts are created based on the date of registration, on a daily basis. (One line is one cohort.) The date of registration is in the first column. The number of users who registered on the given day is in the second column. The rest of the columns show the retention rate of the given cohort calculated within X number of days from registration.

In other words, it's the ((Daily Active Users) / (Registered Users)) ratio within the given cohort: the X-Day-Retention %.
Recommend article: http://data36.com/measuring-retention/

Jun 26th, 2015 – Jul 10th, 2015 DONE

Day | Week | Month

| %

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%

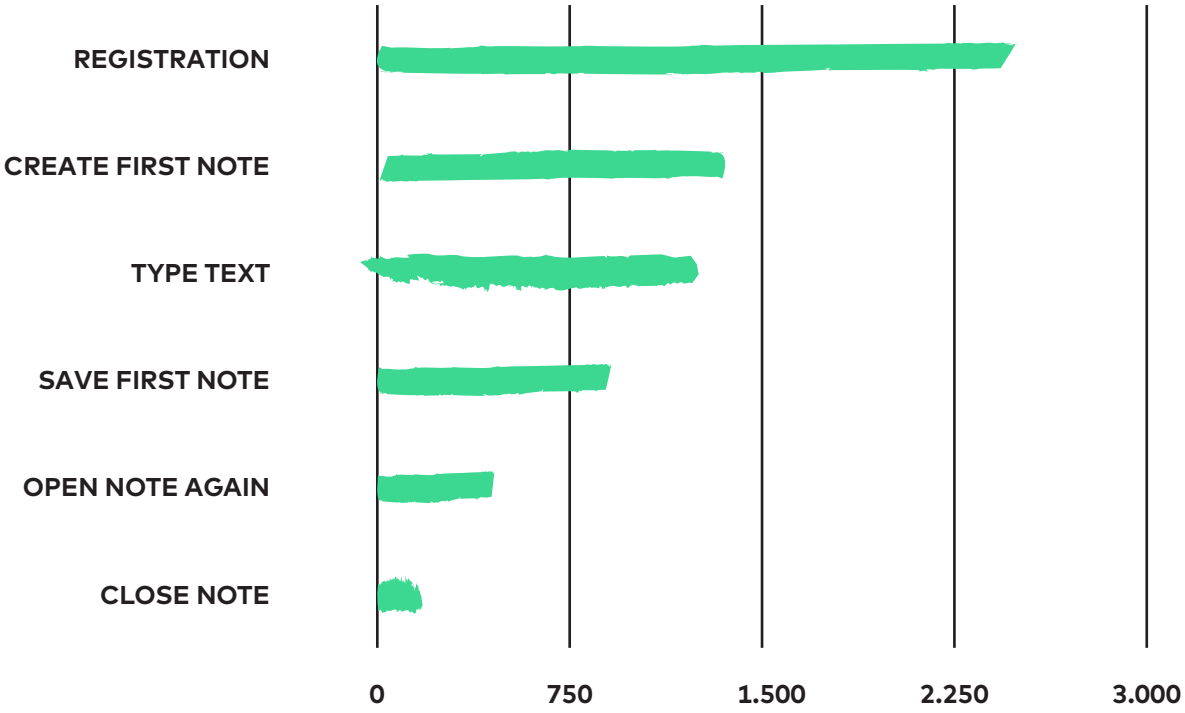
source: Mixpanel

Funnel

Funnel analysis is a powerful analytics method that every online business can take advantage of. It shows the conversion between the most important steps of the user journey. It helps you understand what percent of your users stay with you or churn at a given step. The name “Funnel” comes from the shape of the chart this analysis makes.

Funnel-analysis

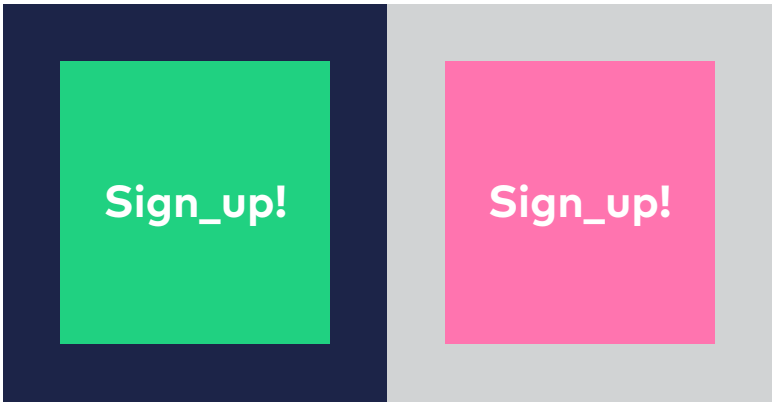
Funnel Analysis is about going through and calculating the conversion between the different steps – to understand how many users reached a certain point of a given process.
The complexity can differ from project to project:
The easiest example is to analyze a registration form. Users fill in these forms from top to bottom. You can expect that fewer and fewer users will fill in each field - they are dropping out step by step.
A well-visualized Funnel looks like this (in this case, for a note app:



Recommended article:
http://data36.com/funnel-analysis/

AB-testing

The testing of two or more alternative versions of any kind of online content. When running an A/B test, every user who lands on the A/B-tested webpage is automatically and randomly put into a test or control group. They see one version of your content. Then you measure what they do there and how many of them reach the assigned goals.



With the right number of users and with some statistical "magic" you can figure out your winning version, which provides the best user experience - and the best conversion rate for you. When you implement an A/B test follow these five basic but important rules:

1. Users should be assigned randomly to test or control groups!
2. Don't let your users know that they are part of a test!
3. The different alternative versions should run at the same time!
4. Define your goals and targets. Make them easy to measure and easy to understand! (E.g. we expect registration rate to grow 20% by the end of the test.)
5. Change one thing at a time!

A common question is what sample size you need for your A/B test. This depends on many things. One is the baseline conversion rate of the control version (e.g. Visit-to-Registration ratio = 3%). The higher this is, the smaller the sample size. The other is the target conversion rate change you aiming for (e.g. you target a 6% Visit-to-Registration ratio - that's a 100% increase). The higher this is, the smaller the sample size. (It might not be intuitive, but check it out!) And finally, the statistical significance you're aiming for (this is usually 95%, but for some tests it's 99%).

Based on all this, Optimizely created a great Sample Size Calculator, which you can access here: https://www.optimizely.com/sample-size-calculator/

Note 1: The traditional example of an A/B test is an e-commerce shop testing a blue "Add to Cart" icon against a red one. (Hint: it's not your most prosperous A/B test...)
But there are many, more complex A/B-tests out there: layout tests, wording tests, title and subtitle tests, creative tests on Facebook, etc...
Note 2: Some articles differentiate the so-called multivariate test from the A/B test. Multivariate testing is the playground of companies with larger user-bases. It works along the same lines as A/B tests. The only difference is that in multivariate testing you can change many things in one experiment - which can be combined with each other in multiple variations of your webpage. There are some advantages of multivariate testing, but if you are new to this topic, start with simple A/B testing instead.

Usability testing

Usability testing is a popular research method for collecting detailed and direct user feedback about your online product, website or mobile application.
So what is usability testing? It's as simple as inviting a user (or potential user) from your (target) audience – then showing her your website, giving her specific instructions and, for around 30 minutes, monitoring what she's doing.
If you do it right, you will get a bunch of useful and actionable insights. You will be surprised how often a button or a description that you consider to be one of the most straightforward things on your website turns out to be the most confusing one for your audience. And that's the point here: to see and feel the actual pain of your users!
Recommended article: https://data36.com/usability-testing

Chapter_06C

A few important metrics to look at

In the previous chapter I described the most often used metrics and analytics methods. But there are many more! Let me give you a few more examples without explaining them in detail.

A few self-evident and extremely useful metrics:

- Avarage 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, go Google them. There are many great articles about them.

And there are quite a few more complex analytics methods to look at. For example:

- Virality Score
- Score Carding
- Regression Analysis
- Clustering
- Principal Component Analysis
- Predictive Analytics and Machine Learning methods
- etc.

Well, there are not enough pages in this mini-booklet to go into detail about all of them.
But here's a recommended article where you can start:
https://data36.com/predictive-analytics

Chapter_07

Case studies

This chapter introduces in short how these concepts are used by real online businesses.

Chapter_07A

E-commerce case study - cohorts and segmentation

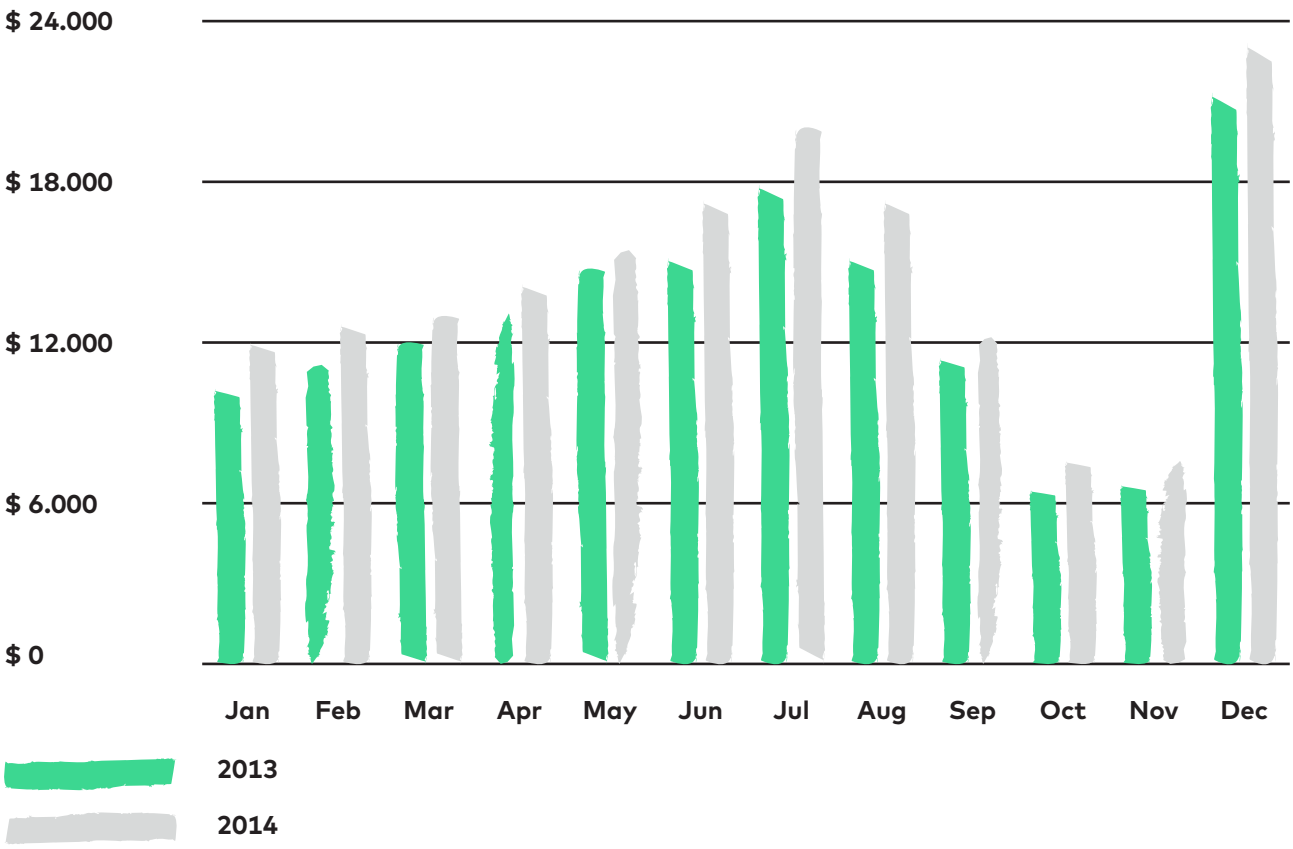
Note: the e-commerce sector is a really tough competitive market, so I had to replace the actual name of the company.

The Hiking Backpack E-Shop (note: this is a fictional name) began to analyze their data. They were curious about:

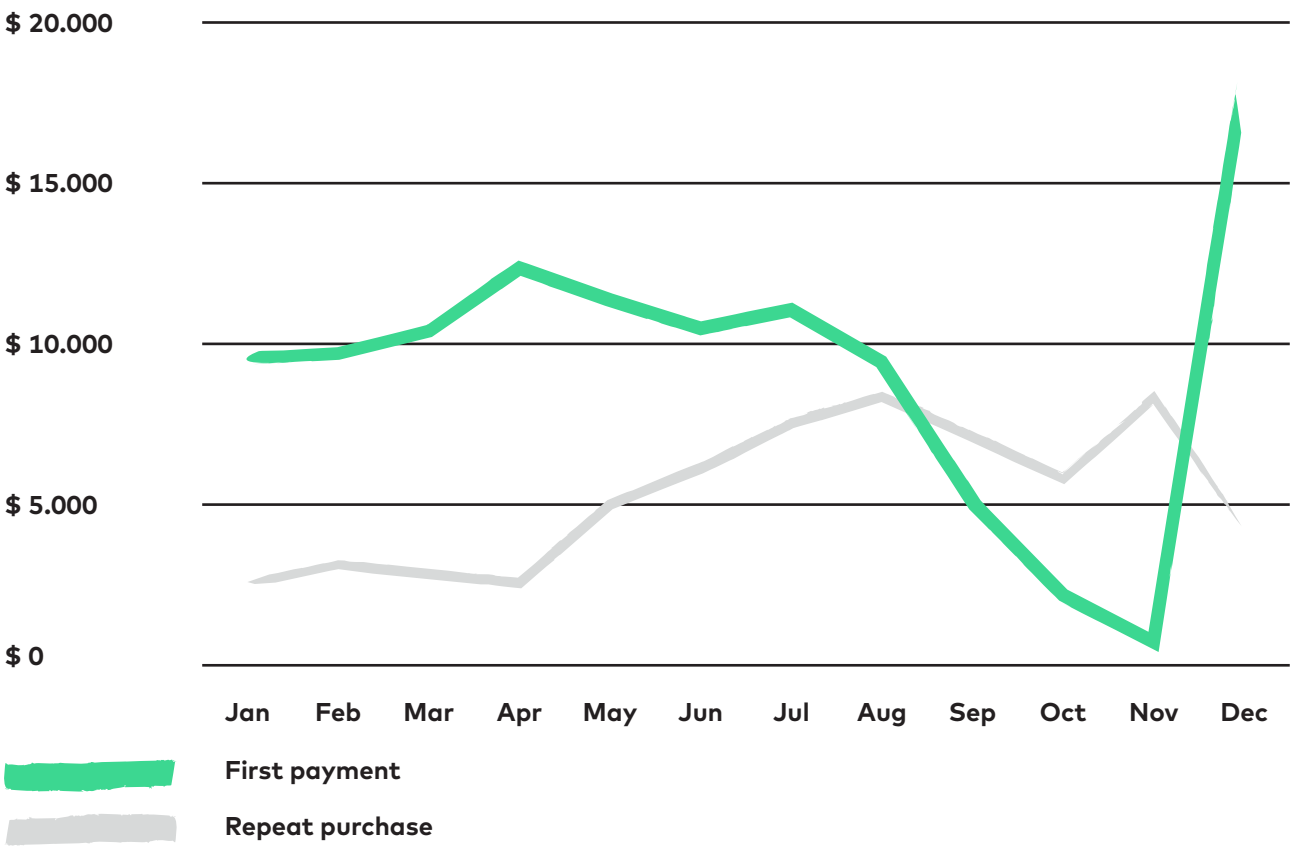
- What is the best target audience for them?
- What kind of product should they offer to whom and when?
- Having answered these two questions, how can they reach the highest revenue and higher customer retention rate in the long term?

The first thing they saw was that sales performance fluctuates throughout the year.

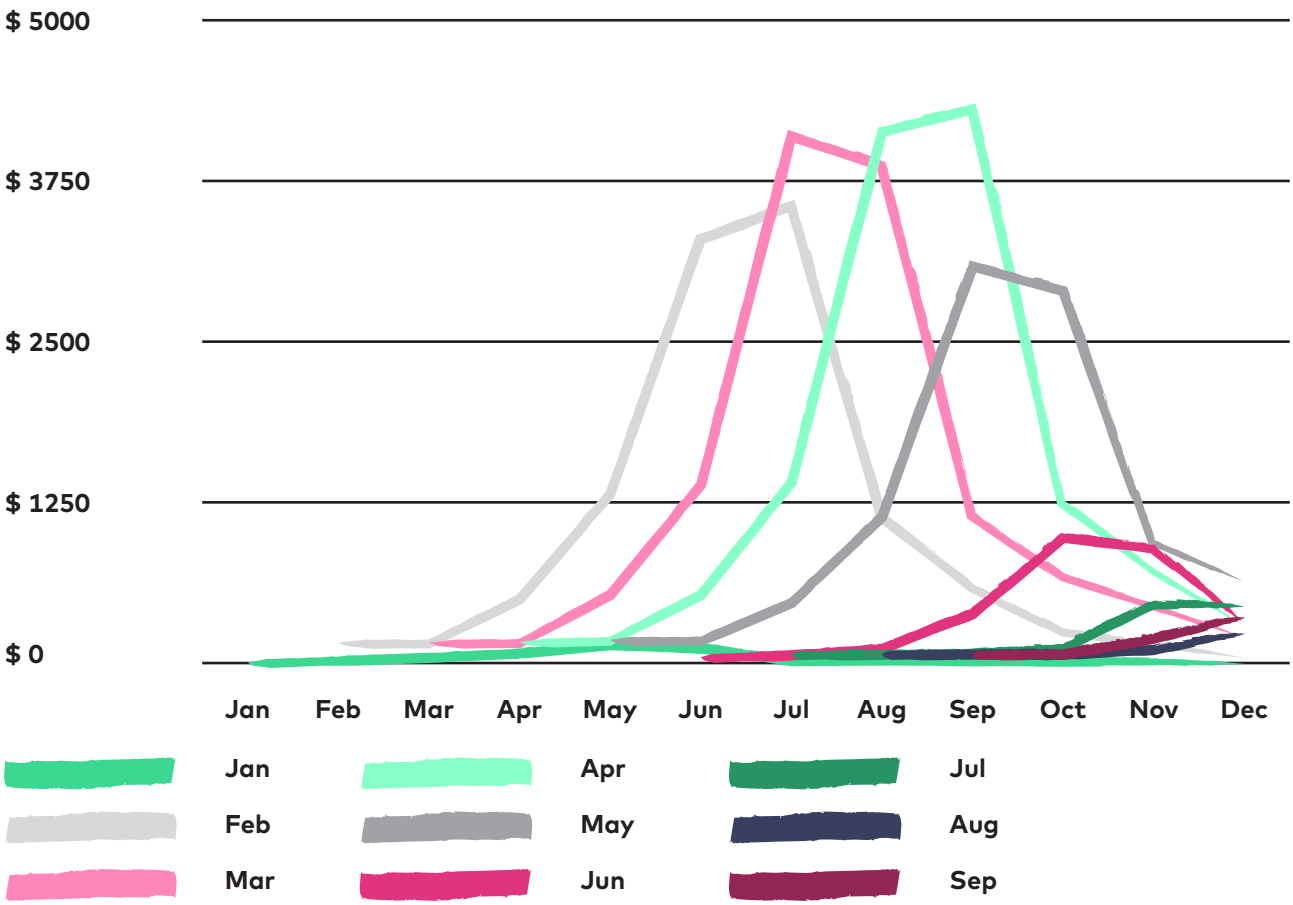
This can be for a number of reasons, of course, but knowing the circumstances, our first thought was that this was due to the nature of the product. (Hiking is seasonal, right?) To validate our suspicions, we looked at the 2013 vs. 2014 revenue chart as a monthly breakdown. The two years show a similar trend (we only see small growth).



We saw the same for 2012 and 2011 as well.
As always, we did a number of user interviews and usability tests, and checked some obvious analyses based on different hypotheses. Most of these didn't give us any useful information – but one of the segmentations had an interesting result.
We segmented the revenue on the below chart based on purchase history.

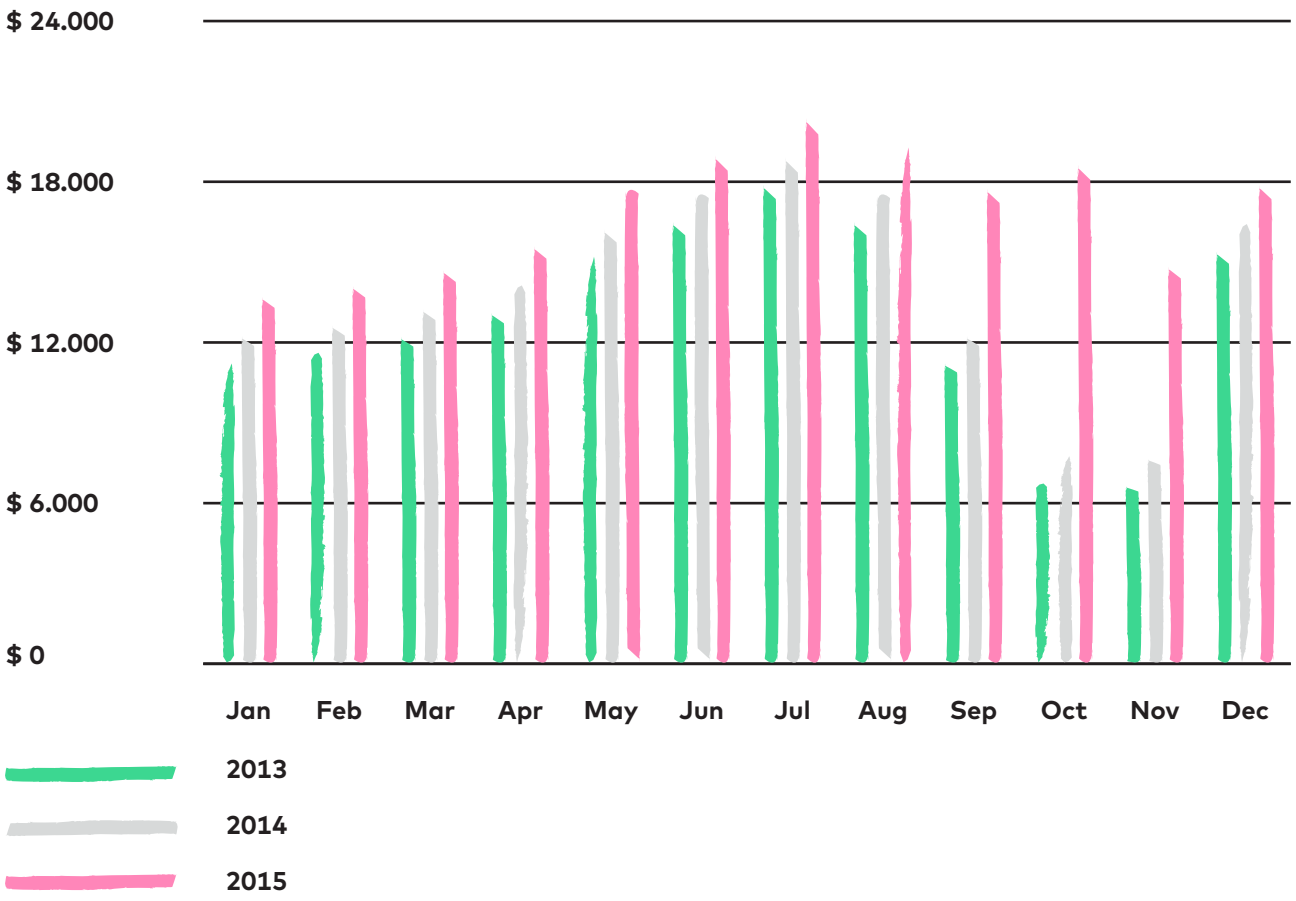


We can see that there was a constant change in 2014 on whether the "simple" first payments (the first purchase of a given customer) or the repeat purchase (when a returning customer purchased again) brought in more revenue.
It jumps out that although revenue generated by new customers drops in autumn, 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 the first time – and then second, third, fourth (and so on) times that they purchased something. We found this:



In 2014, customers spent the most on repeat purchases at the end of the summer and at the beginning of autumn. In fact, we also know that customers from February, March, April and May are really loyal and spend a lot 4-5 months after they make their first purchase (so in July, August, September and October).

From this, two obvious suggestions followed.
One was to take a look at the same metrics, but over many years. (These also showed that customers who first bought gear between February and May came back and spent much more later in the year. We slowly realized that this means that they were the cohort who took hiking more seriously, planned their trips ahead of time, and bought their gear in advance. The rest shopped on an ad-hoc basis in the summer, or gave hiking backpacks as gifts, etc.)
The other was to figure out what other products can be sold to the loyal customers. This was a much simpler story. In short – we were able to find an easy-to-target customer group and also figure out what and when to sell to them again.
The autumn campaign of 2015 was approached with a brand new retention strategy. Instead of aiming at new customers, the company focused on pre-existing ones during these 3 months. You can see the results on the graph:


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

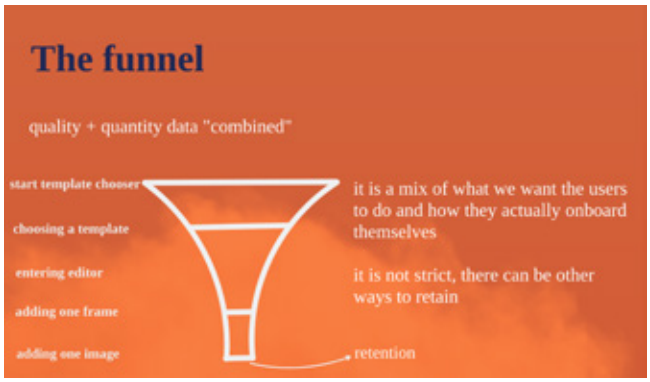
Chapter_07B

Funnel analysis at Prezi

Andris Balogh is the former senior Lead Data Analyst at Prezi. During his time there he gave an insightful presentation on how and for what he uses Funnel analysis. *Note: This was at the BData 2014 conference (organized by Data36).*

"[...] When we have collected all the information from the analyses and have sat down with the Social Researcher and UX Researcher, we think over what kind of *Funnels* a User has 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. At least, this is the structure which has come out of the analyses, usability tests and all other types of research.

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. In a reg form you provide your name, email address, click on the registration button, click OK and then you are done... In comparison: with the Prezi Editor a *User* can take many, many paths.



Due to this, a Funnel is a mix of what we want the users to do (based on how he/she understands the product), as well as what the users actually do based on the analyses.

So it's an interesting synthesis between expectation and reality. It's important to see that since this is not a linear *funnel*, the *user* can come back in different ways. Regardless, we still created a funnel which mainly captures those who continuously stay *active users*. And those who 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, either. I think the best option is if we begin our work at the bottom of the funnel. Because if you begin to pack users to the top of the funnel (e.g. with Google AdWords or Facebook Ads), those will churn 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 you know love your product and have tried many of your 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 with larger marketing costs and bringing in more users.

Let's look specifically at the case of Prezi. In this case, inserting the first image was the most important funnel step. This is a real decision of a user: creating presentations begins with inserting an image... at least at Prezi.

Thus the Developer, the UX Researcher and the Designer sit down and begin to work around this function. During this, there are ongoing smaller analyses of course, as it's better to pinpoint the exact problem with images that we want to fix. Usability Tests can also change into something that only deals with image placement. The analysts can create a "higher resolution" subfunnel for images.
For instance:

1. User press the „add image" button, then
2. User clicks 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 A/B testing works at their company. Here's a short extract:

"One of our projects is about explaining to the users the easiest way to purchase Ustream Pro Broadcasting and what kind of extra features they will have.
(One of the most important features is that you can broadcast to your viewers advertisement-free.)

To go Pro, users have to fill in a registration form where we asked pretty much everything about them. We tried to optimize this form so we have as many subscribers as possible.
This A/B test was about whether to have an overview page where Users can check the 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 here comes the interesting thing! We noticed much later – which was not even measured in the original A/B test – that the number of Refunds differed. Those who received the overview form requested a Refund much less often than those who received the shortened form... So we only realized way after the experiment that from this perspective, the overview form version was a clear winner. We understood that it's important to follow up on every A/B test 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 how "codeless testing" works and why it's worth doing. Here is one of the most useful parts 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 usually 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. Before the test, 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 to read this mini-booklet. I know data science is not a simple topic, and a data dictionary especially - with all its definitions - may feel like a dry read. But it's so important!

I hope you can make practical use of what you read here -- and create a consistent and thought-out common language of data science in your organization. As I mentioned in the intro chapter, the goal is not to dictate a 100% match with what is written here, but 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 at this email address:

tomimester@data36.com

Check out my workshops for companies:

https://data36.com/data-science-company-workshops/

And my online courses:

https://courses.data36.com

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!

Design by Faraway Design