

BigSurv18

Learning on Survey Data to Qualify Big Data in a
Web Environment

06/11/2018







Mediametrie



Our Company



 Founded in 1985	 670 employees	 > 1 000 customers in France and worldwide	 Turnover 100 M€
---	---	--	---

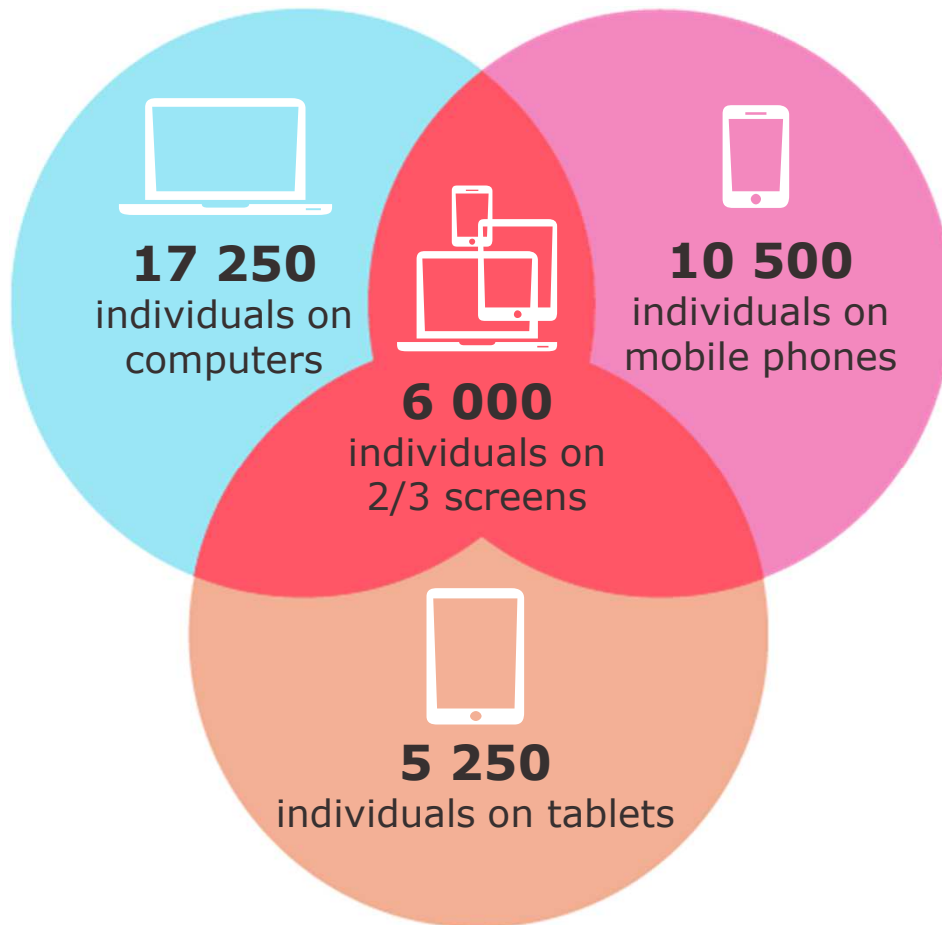
-  Leader of Media and reference studies in audience measurement
-  Methodological and technological innovations
-  Tracks the changes in the Media ecosystem
-  Sheds light on public behaviour

Mediametrie is the benchmark in audience measurement in France for the TV, the radio and the Internet




**Panel for the Internet audience, global and per screen (computer, mobile, tablet)
30 000 panelists**

Internet audience measurement



3 panels for a Global Internet measurement :

Computers

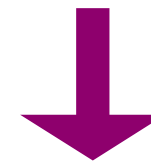
- People aged 2 and over
- Based on Nielsen meter

Mobile phones

- People aged 11 and over
- Based on RealityMine meter

Tablets

- People aged 2 and over
- Based on proxy and survey app



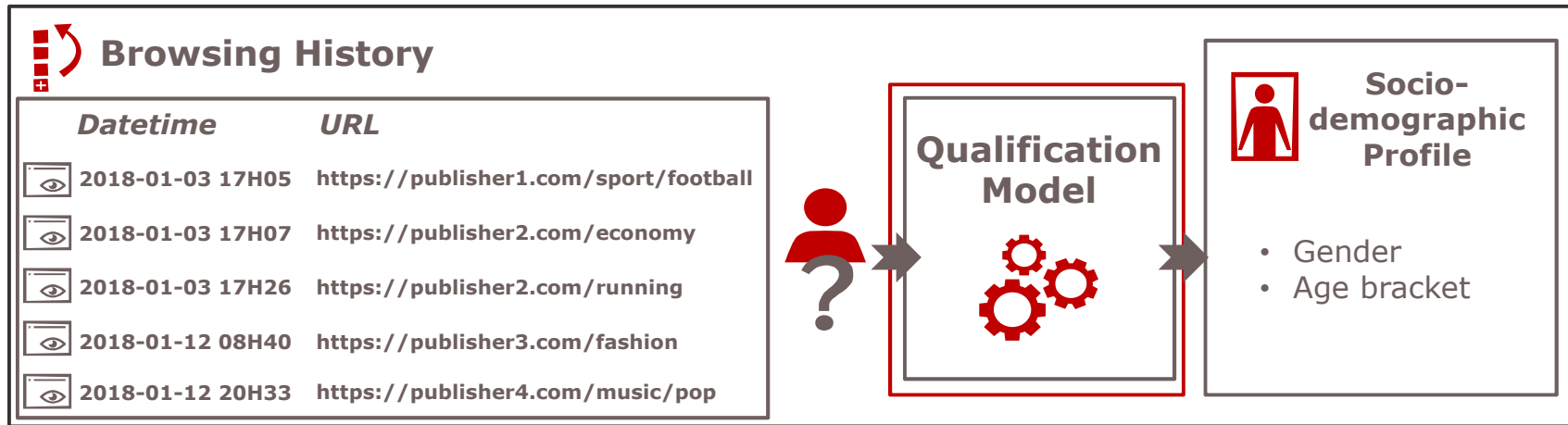
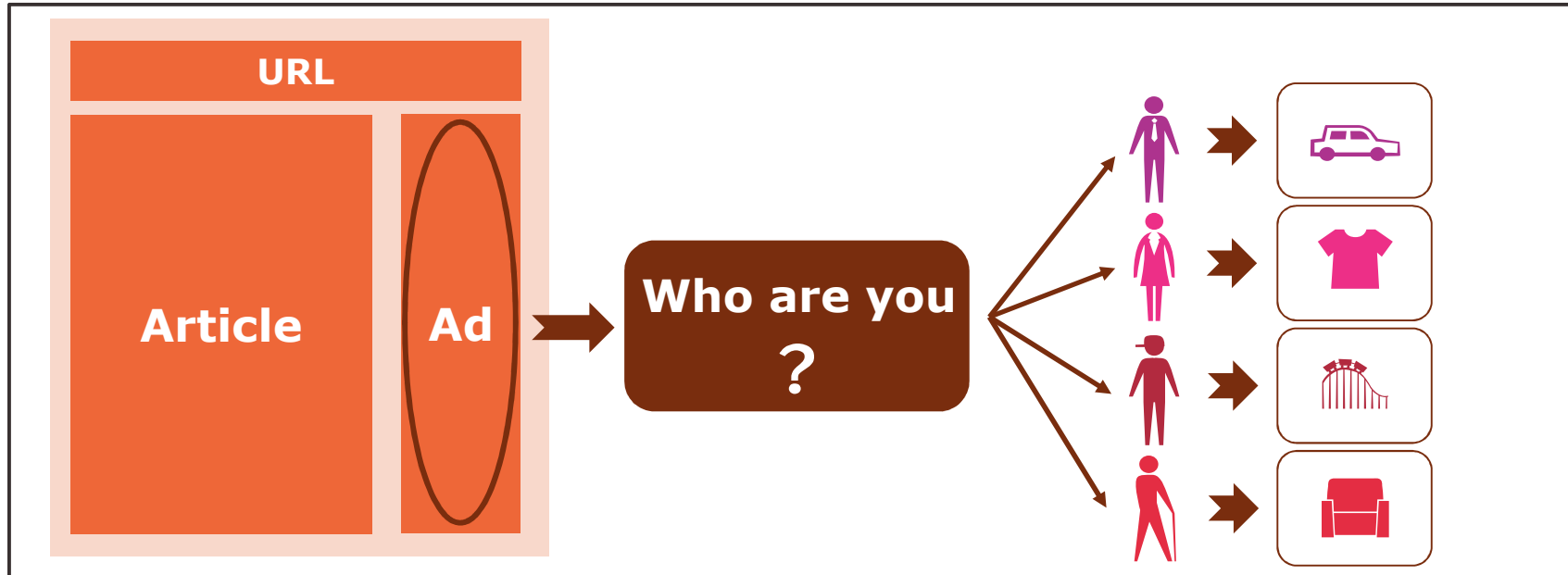
Global Internet measurement

- Statistical fusion of the 3 panels
- Metrics based on single source sub-sample
- Weighting process including an adjustment on site-centric figures

Customer Context



Data Profiling Web



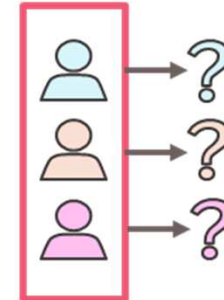
Data



Client Data

Data to qualify

cookie ID	datetime	URL
123456789	2018-01-03 17H05	https://publisher1.com/sport/football
123456789	2018-01-03 17H07	https://publisher2.com/economy
123456789	2018-01-03 17H26	https://publisher2.com/running
104785236	2018-01-03 19H22	https://publisher2.com/politics/election-results
369874562	2018-01-03 19H43	https://publisher1.com/sport/basketball/Video/How to shoot a basketball with power and accuracy
369874562	2018-01-03 21H31	https://publisher3.com/fashion/Video/fashion week
104785236	2018-01-05 02h48	https://publisher2.com/economy
104785236	2018-01-05 03h09	https://publisher2.com/economy/France
123456789	2018-01-12 08H40	https://publisher3.com/fashion
123456789	2018-01-12 20H33	https://publisher4.com/music/pop
369874562	2018-01-12 20H33	https://publisher2.com/news/
369874562	2018-01-15 00H06	https://publisher2.com/news/~today 2018 01 15%/
369874562	2018-01-15 15H25	https://publisher2.com/globaleconomy/business/start-up
369874562	2018-01-16 15H26	https://publisher2.com/globaleconomy/business/start-up2



Panel Data

Training Data

computer ID	browser ID	datetime	URL	panelist ID	age	gender
KJH	browserA	2018-01-03 17H05	https://publisher1.com/sport/football	123	44	Male
KJH	browserA	2018-01-03 17H07	https://publisher2.com/economy	123	44	Male
KJH	browserA	2018-01-03 17H26	https://publisher2.com/running	123	44	Male
PNG	browserA	2018-01-03 19H22	https://publisher2.com/politics/election-results	456	27	Female
DFY	browserB	2018-01-03 19H43	https://publisher1.com/sport/basketball/Video/How to shoot a basketball with power and accuracy	789	35	Male
DFY	browserB	2018-01-03 21H31	https://publisher3.com/fashion/Video/fashion week	789	35	Male
PNG	browserA	2018-01-05 02h48	https://publisher2.com/economy	951	33	Male
PNG	browserA	2018-01-05 03h09	https://publisher2.com/economy/France	456	27	Female
KJH	browserC	2018-01-12 08H40	https://publisher3.com/fashion	123	44	Male
KJH	browserC	2018-01-12 20H33	https://publisher4.com/music/pop	123	44	Male
DFY	browserB	2018-01-12 20H33	https://publisher2.com/news/	789	35	Male
DFY	browserB	2018-01-15 00H06	https://publisher2.com/news/~today 2018 01 15%/	357	38	Female
DFY	browserB	2018-01-15 15H25	https://publisher2.com/globaleconomy/business/start-up	789	35	Male
DFY	browserB	2018-01-16 15H26	https://publisher2.com/globaleconomy/business/start-up2	789	35	Male



Feature Engineering

Python code 



- **Datetime : when does the cookie visit the perimeter ?**



- **Days** of week
- **Time** slots
- Days + Time slots



- **URL : what sort of content the cookie is interested in ?**



- **Keywords** contained in URLs (music, sport, fashion,...)
- **Domain** names relative to the URL
(*www.domain.com/article/title_of_the_article*)



- **Semantic analysis** of the URLs
 - Preprocessing with tokenization, stop words removal and French stemmer
 - Word clustering
 - URL clustering



- **Surfing : what is the cookie behaviour when surfing on the perimeter ?**

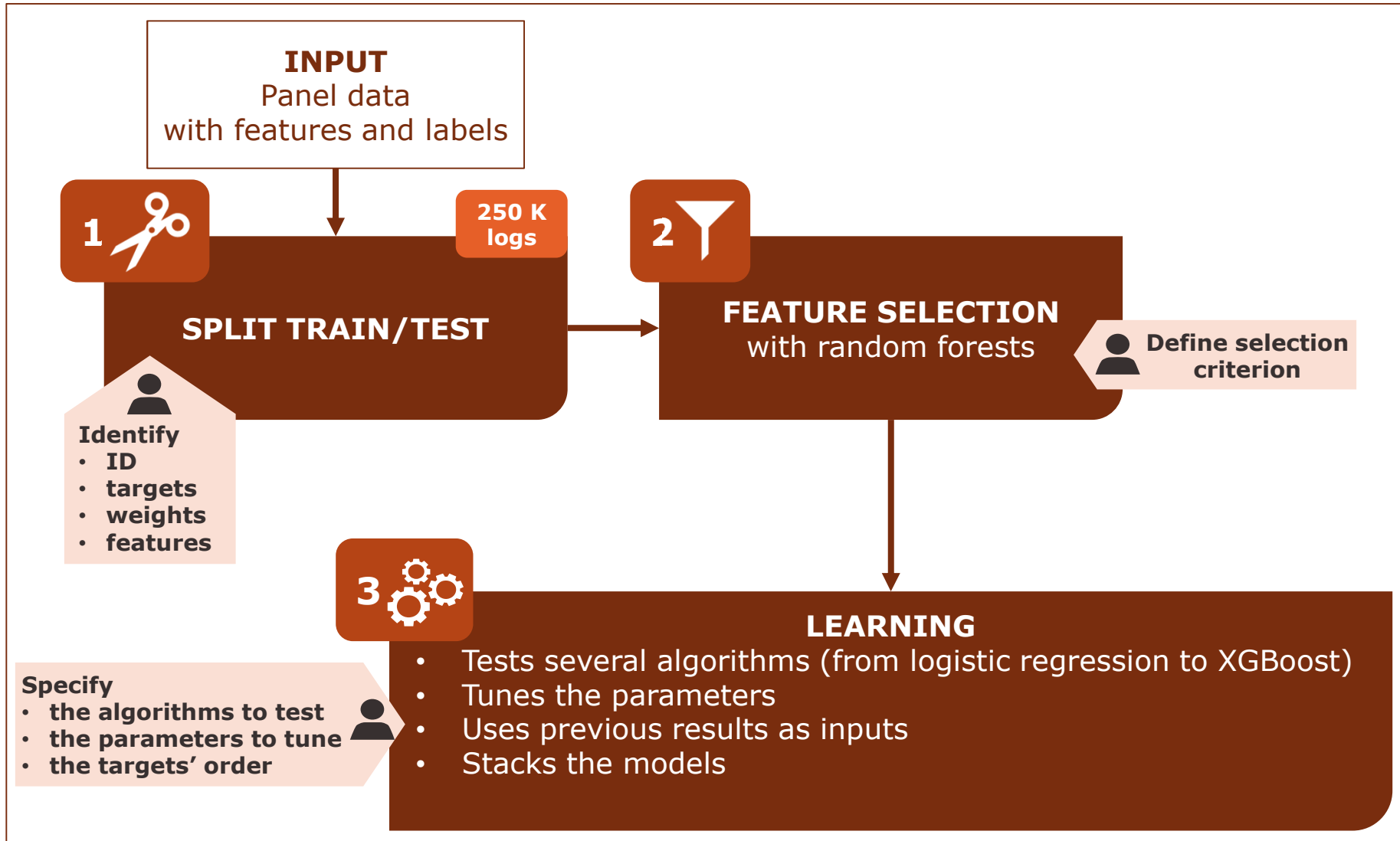
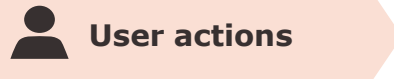


- **Time between** 2 URLs
- Number of URLs per **session**
- Number of URLs per **day**
- **Consecutive days** without logs



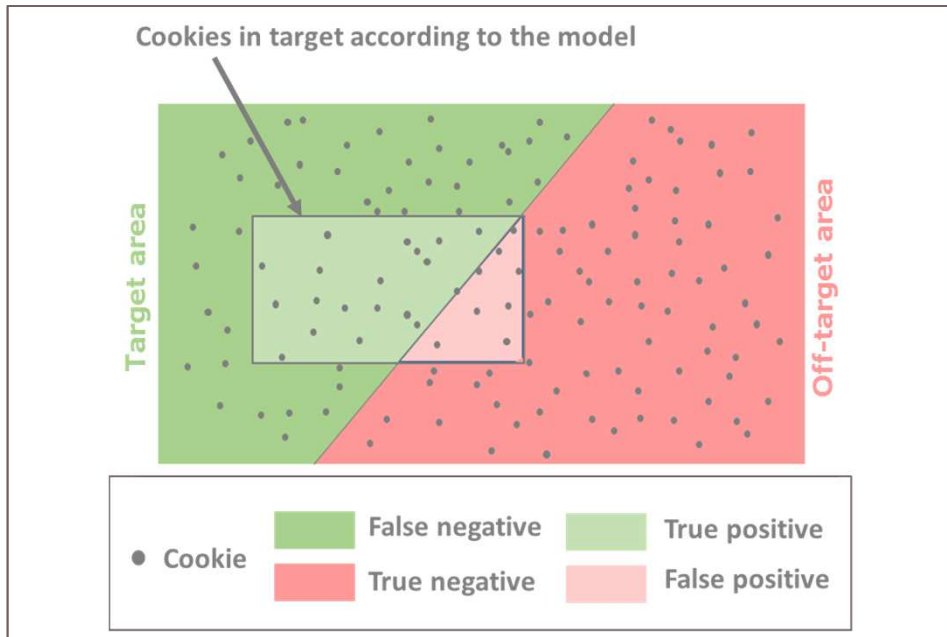
~1300 features

Qualification : Workflow





Results analysis

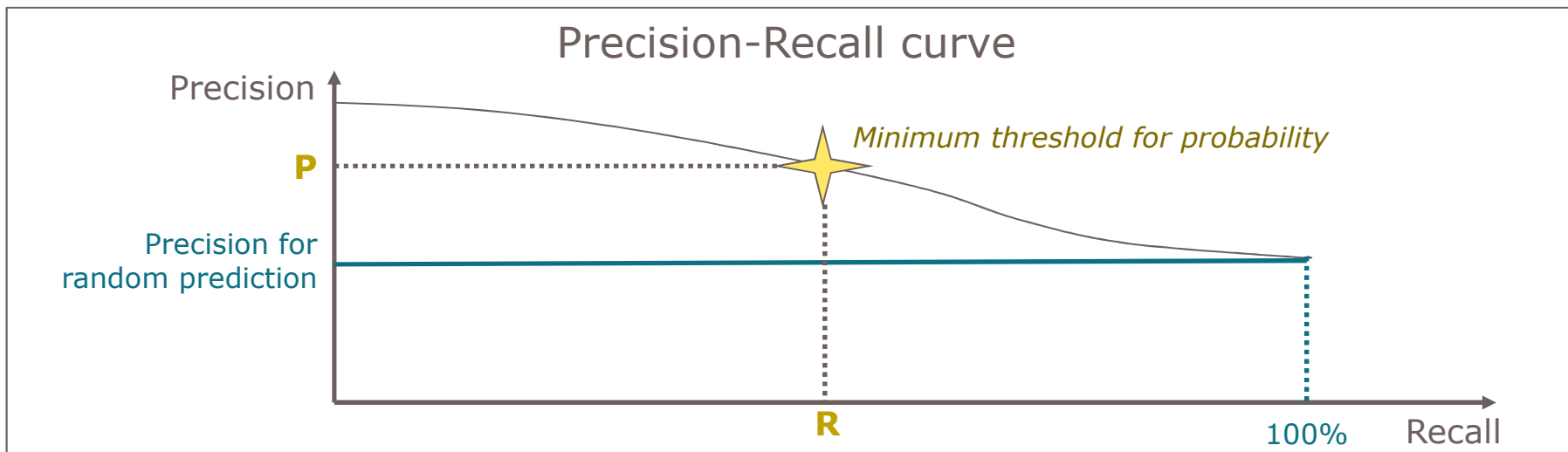


For the cookies **predicted in** the target, how many of them are really in ?

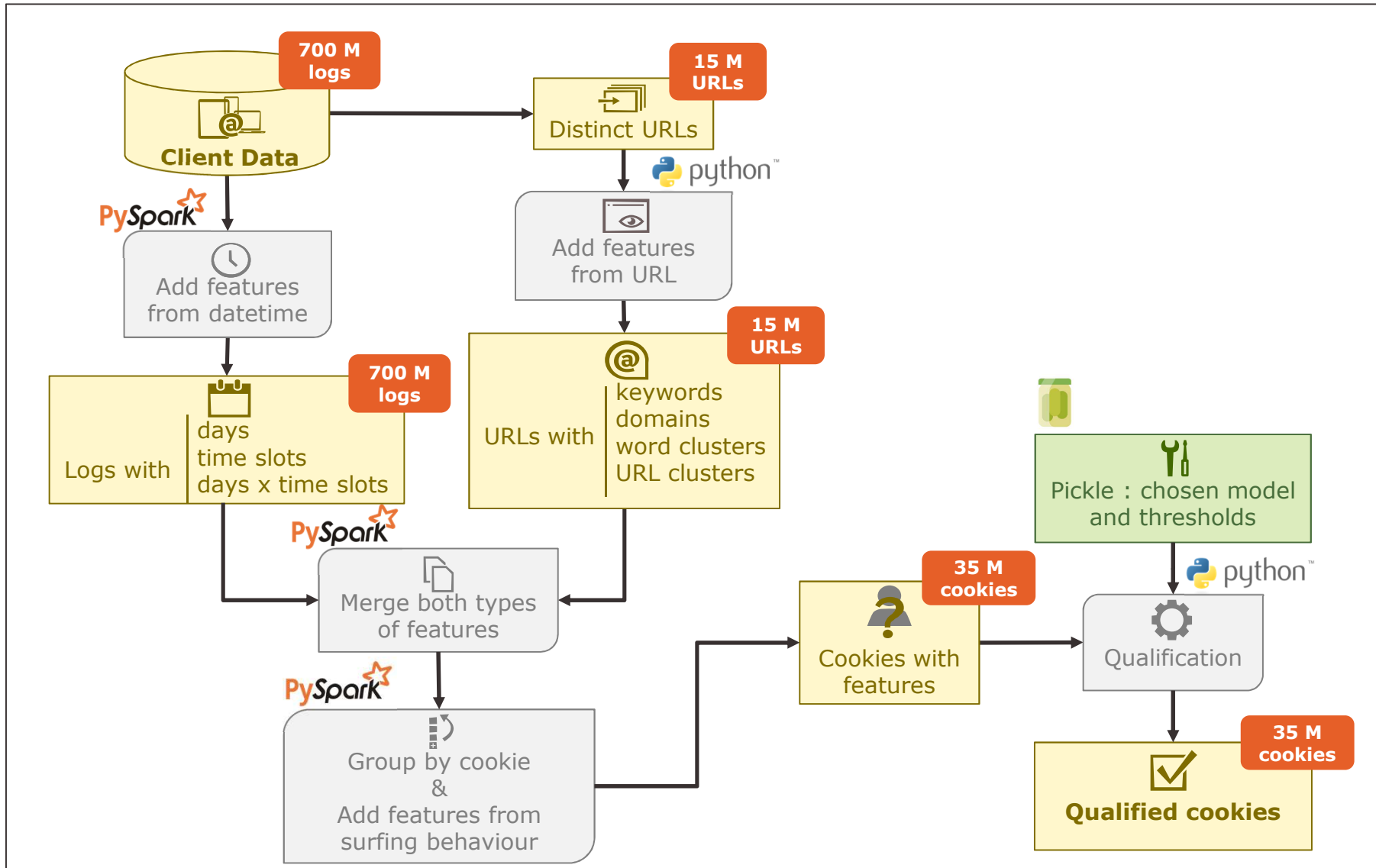
$$Precision = \frac{True\ positive}{True\ positive + False\ positive}$$

For the cookies **in** the target, how many are found by the model ?

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}$$



Project scaling







Conclusion

Today

- **Checking the qualification performances in real conditions**
 - Real advertising campaigns
 - Results monitoring
 - Tests on different targets



Next steps

- **Becoming the qualification partner**
 - Qualifying in a production process
- **Improving the model**
 - Adding new features based on Graph Theory
 - Research 
 - Updating 
 - Testing other algorithms

