The headache of data integration

Clément Levallois

2017-31-07

Table of Contents

last modified: 2018-05-18

1. Data: you don’t get in on tap

A naive vision of data would get that it’s all fluid. Don’t we talk about "data streams"? Working with data would be as simple as opening a tap and "getting customer data", for instance.

Figure 1. Data streams, as fluid as water on tap?

Actually, data is more like a complex patchwork: many different pieces which must be stitched together - and this is hard.

Figure 2. Data streams make a patchwork

Take customer data.

It is not a given. Instead, this is a design made of multiple primary data sources:

Personal	Preferences	Permissions, exclusions, interests
	Demographics	Volunteered & inferred lifestyle attributes
Marketing Treatments	Promo History	Offers by type, product, channel, date & campaign
	Response History	Responses by offer, channel, date
Behavior	Interactions	Summary of sales & service contacts, dispositions
	Transactions	Purchases, reservations, cancellations, changes
Predictions	Planned Promos	Candidate & optimized offers from recent analytics
	Model Scores	Propensity & value estimates

Source: UNICA Corporation, in *Multichannel Marketing*, by A. Arikan (2008).

Figure 3. Multiple sources of customer data

> Analysts often spend 50-80% of their time preparing and transforming data sets before they begin more formal analysis work.

Garrett Grolemund, Data Scientist and Master Instructor at RStudio

Take away: Data is fragmented by nature. It comes from different sources and presented in different formats. **You** (the marketer in collaboration with data scientists) wrangle to *construct* customer profiles by joining and assembling different sources of data into a meaningful synthesis.

Data scientists actually have entire books devoted to the subject of wrangling with the complexity of integrating different data sources.

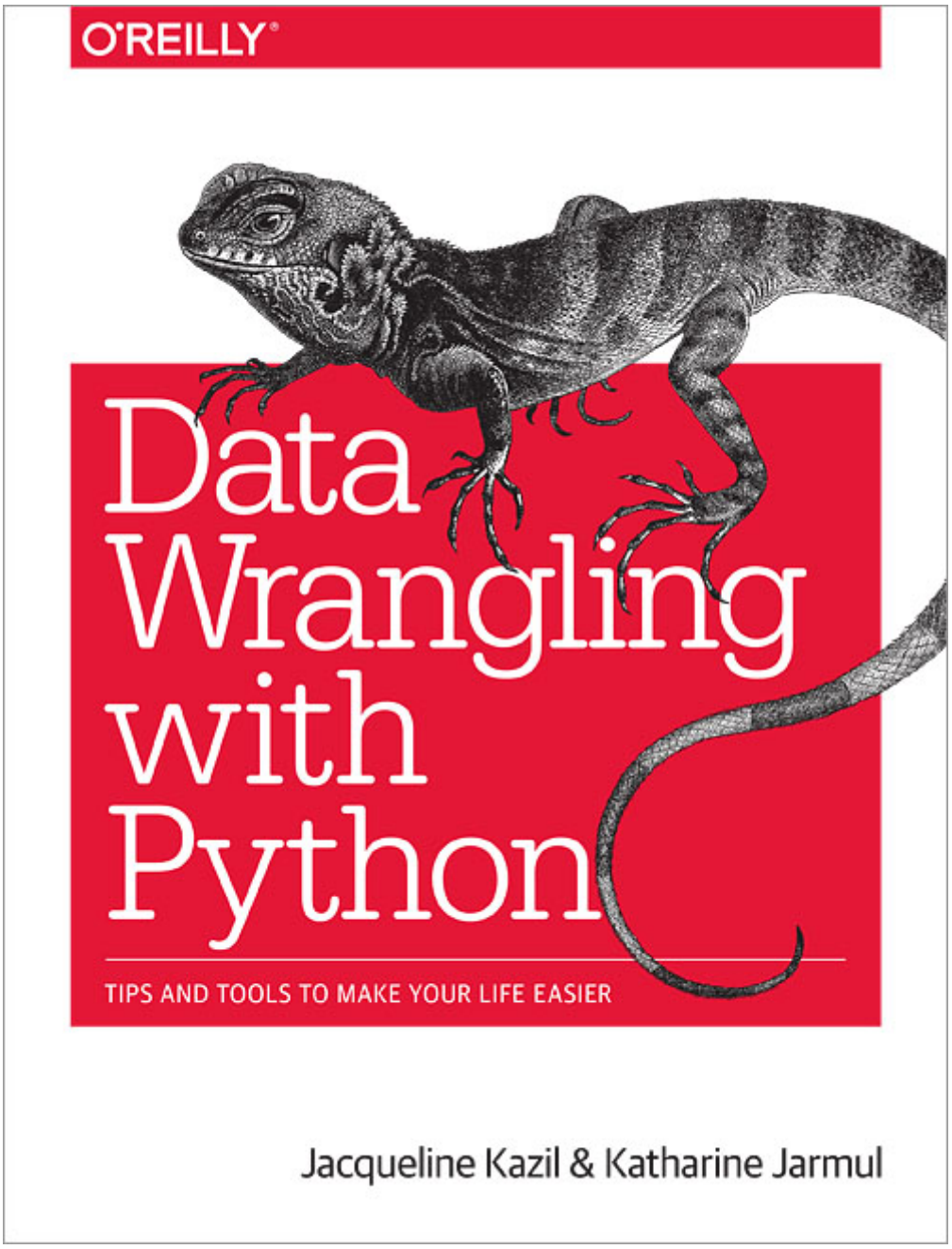

Figure 4. Data wrangling

2. Sources of fragmentation

a. Channels keep diversifying

Point of sale, print, TV, radio, outdoor posters, mobile apps, mobile sites, emails, SMS, APIs, social networks, search engines, e-commerce platforms, e-commerce websites, blogs, content channels, ...

→ all these channels can provide relevant data.

b. Connections between these channels intensify and complexify

- Social TV is TV delivered with Internet services,
- User profiles created on one platform are imported on another
- Orders taken online can be picked up on a variety of point of sales
- Ads circulating through one channel replicate on other channels, ...

→ It is very complex to trace the "customer journey" on all these channels and to keep an updated view of a customer profile.

It is even more difficult to explore causality (which action on which channel caused which subsequent action by the user?)

c. Underlying technologies fragment and keep evolving, across channels

Browsers, Cookies, APIs, mobile OS (Android or iOS?), etc... All these different techs evolve and need continuous effort and expertise to integrate.

Example: **did you notice** that on a mobile device, the url of the pages you visit can now start with an AMP url?

Like, to visit the page of the New York Times the url should be http://www.nyt.com but it looks like: **http://google.com/amp/www.nyt.com**

This **http://google.com/amp** prefix is a new tech by Google to accelerate the display of web pages on mobiles. Fine. But then, as a marketing data analyst, how to count visits to:

http://google.com/amp/www.nyt.com

and

http://www.nyt.com

→ It is important to count visits to these two urls as a visit **to the same page**.

In September 2017 major services of web analytics were still struggling with this issue of page counts in web analytics.

This illustrates that to just count visits to a web page (something which should be classic and robust) and integrate this data to a larger analysis, big issues can arise and be hard to fix even in 2017, because of the evolution of techs and standards.

d. In the meantime, customers have growing expectations about the quality of service

Difficulties posed by data integration do not slow or decrease customer expectations. To the contrary, we see an elevation of expectations. Customers increasingly expect:

- realtime contact
- two-ways interaction (they want to be able to voice their opinion, and get a response)
- seamless experience (no glitch, modern user interface, consistence of the user experience across channels)
- personalized experience (customization of the message they receive)

e. Example: A French bank going through the 2010s

Figure 5. Before - a couple of data sources across a few channels

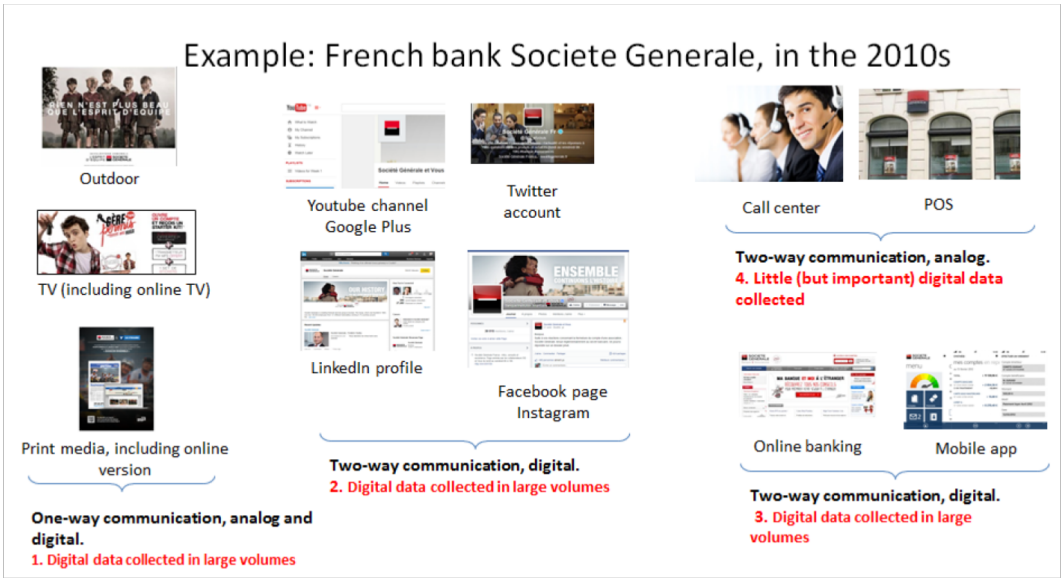

Figure 6. Now - many data sources across a variety of channels

3. Tools for data integration: DMPs and more

a. Data Management Platform (DMP)

In 2015/2016 a new acronym started to trend: "DMP", standing for **"Data Management Platform"**.

Basically a **DMP** is an information system dedicated to solving the issues of data integration:

- it can store a large amount of data
- it can receive data from a variety of sources, in a variety of formats
- it offers functions to reconcile records from different data sources and generate a unique identifier for each reconciled entry.
- it offers segmentation / classification functions
- it provides security and analytics capabilities on the data
- it makes this data available for execution by other software.

b. DMP in relation to other components of the information system

DMPs are relatively new. They integrate with 3 other information systems in the firm:

- CRM
  - This is the software **gathering** data related to customers and sales. It is a major source of **input data** for a DMP.

- ERP
  - Large software synchronizing information systems from finance, sales, logistics and more. The CRM can be independent or part of the ERP.
- DSP
  - Piece of software automatizing ad buying. The audiences identified in the DMP could be served corresponding ads with a DSP.
- Data lake
  - Data lakes are databases specializing in storing large amounts of unstructured data. They respond to the need of "storing today, for future use". A data lake is not meant to be optimized for queries. They are meant to store everything we collect today, in the case this data will serve future usages. When this happens, data can be extracted from the data lake and put in a database, in a form which makes it convenient to query and analyze.

How can data circulate across these software and with the external world? The next chapter is devoted to APIs, another essential concept.

The end

Find references for this lesson, and other lessons, here.

This course is made by Clement Levallois.

Discover my other courses in data / tech for business: https://www.clementlevallois.net

Or get in touch via Twitter: @seinecle