# Open Source in AI

Author, Francesco Corea

A Data Science Foundation Blog

January 2019

---------------------------------------------------

www.datascience.foundation

*Why it matters to give away your software for free*

## I. Introduction

Open sourcing some technologies is kind of counterintuitive at a first look. Why on earth should a company give away what they invested money and people in? I have already written on this trend, but I keep sharpening my thinking around it and this post is a consequence of recent new considerations.

The **open source model** is quite hard to be reconciled with the traditional SaaS model, especially in the financial sector. However, we are observing many firms providing cutting-edge technologies and algorithms for free. While in some cases there is a specific business motivation behind it (e.g., Google releasing Tensorflow to avoid conflict of interests with their cloud offering), the decision of open sourcing (part of) the technology actually represents an emerging trend.

Tools are nowadays less relevant than people or data and the sharing mindset is a key asset for organizations. Based on this statement, we can divide the considerations on open source in two clusters, which are ***business considerations*** and ***individual considerations***.

## II. The Business Perspective

From a business perspective, the basic idea is that is really hard to keep the pace with the current technological development and **you don't want your technology to become obsolete** in three months time. It is better to give it out for free and set the benchmark rather than keeping it proprietary and discard it after a few months. Furthermore, open sourcing:

- Raises the bar of the current state of the art for potential competitors in the field;
- Creates a competitive advantage in data creation/collection, in attracting talents (because of higher **technical branding**), and creating additive software/packages/products based on that underlying technology;
- Drives progress and innovation in *foundational technologies* (thanks Wes McKinney for pointing out this aspect);
- Increases the overall *value, easy of integration* and *reliability* of internal closed source systems;
- Raises awareness of the problems faced at scale on real-world data;
- Lowers the adoption barrier to entry, and gets traction on products that would not have it otherwise;
- Shortens the product cycle, because from the moment a technical paper is published or a software release it takes weeks to have augmentations of that product;
- More importantly, it can generate a *data network effect*, i.e., a situation in which more (final or intermediate) users create more data using the software, which in turn make the algorithms smarter, then the product better, and eventually attract more users.

## III. The Individual Perspective

From the developer's point of view instead, there are a series of different interesting considerations:

- Github accounts and packages look better and have a greater impact than a well-written resume in this world;
- Data scientists and developers are first of all scientists with a sharing mindset, and part of the industry power to attract and retain talents come from augmenting the academic offer (i.e., better datasets, interesting problem, better compensation packages, intellectual freedom);
- Academia has been drained of talents who moved to the industry and the concept of '*academic publication review*' has been translated into '*peer review*' (***crowd-reviewing***). This is in turn translated into i) **better troubleshooting**, ii) deeper understanding of technology potential and implications;
- ***Making codes that others can read and understand is what makes you better developer and scientist.*** This is something you know only if you have ever done it;
- As a general rule-of-thumb, if the contributors are from academia, they usually push innovation forward, while industry contributors prefer more system stability. Releasing open source software helps you thinking about who will use it and design the entire software more reliable and stable in the first place.

### IV. Effects of open source model on AI development

These are some of the reasons why this model is working nowadays, even though there are advocates who claim incumbents **to not really be maximally open** (Bostrom, 2016) and to only release technology somehow old to them.

My personal view is that companies are getting the best out of spreading their technologies around without paying any costs and any counter effect: they still have unique large datasets, platform, and huge investments capacity that would allow only them **to scale up**.

Regardless the real reasons behind this strategy, the effect of this business model on the AI development is controversial. According to Bostrom (2016), in the short term, a higher openness could increase the diffusion of AI. **Software and knowledge are non-rival goods**, and this would enable more people to use, build on top of previous applications and technologies at a low marginal cost, and fix bugs. There would be strong brand implications for companies too.

*Software and knowledge are non-rival goods.*

In the long term, though, we might observe **less incentive to invest in research and development**, because of free riding. Hence, there should exist a way to earn monopoly rents from ideas individuals generate. On other side, what stands on the positive side is that open research is implemented to build absorptive capacity (i.e., it is a mean of building skills and keeping up with state of art); it might bring to extra profit from owning complementary assets whose value is increased by new technologies or ideas; and finally, it is going to be fostered by individuals who want to demonstrate their skills, build their reputation, and eventually increase their market value.

### V. The war of talents

I am adding a final concept which I find intriguing but a bit speculative. It is about the ***war of talents*** and the uncanny vicious circle we are observing in the academia-industry relationship.

The problem is indeed twofold:

1) Universities are ***losing faculty and researchers to the benefit of private companies***. This does not allow universities to train the next generation of PhD students which are now driving the AI wave;

2) Things are already moving and many graduate students are deciding to not go for a PhD at all and jump directly into the private tech sector. This means that **we might not have at all a new generation of PhDs.**

So no students, no teachers. What have we left? **Training on the job.** I believe open source is helping private companies to gradually be recognized as new ‘***knowledge labs’***—they already were in my opinion, but traditionally this role has been assigned to universities.

It is an additional way in which big incumbents are taking over universities with a new indirect approach to education which might eventually disrupt the traditional learning we all know.

References

Bostrom, N. (2016). “Strategic Implications of Openness in AI Development”. Working paper.

Note: the above is an adapted excerpt from my book “Big Data Analytics: A Management Perspective” (Springer, 2016). A new version of this article has been proposed in “Introduction to Data” (Springer, 2019).

# About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

# Contact Data Science Foundation

Email: contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation