



Data scientists vs. Statisticians: Lessons learned after two years of practical experience

Prof. Dr. Bertrand Loison, Swiss Federal Statistical Office, Vice-Director

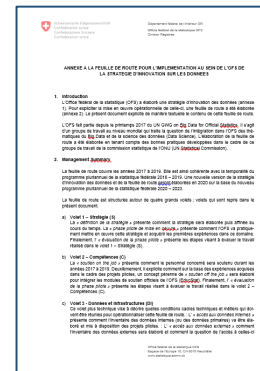
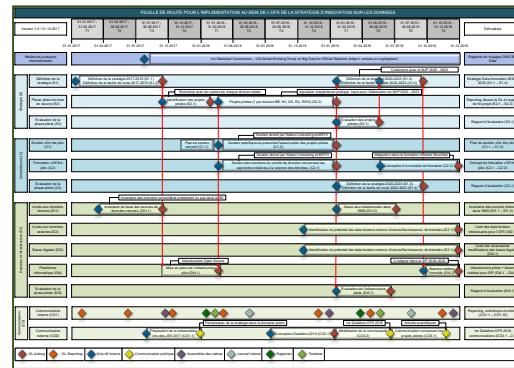
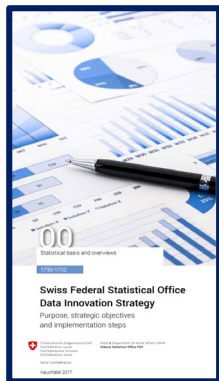
5th International Conference on Big Data for Official Statistics

Session “Data Science and Capacity Development in Official Statistics”

Kigali, Rwanda, 3 Mai 2019



Two years of practical experiences (2017 – 2019)



Fso's Experimental statistics

Experimental statistics are produced using new methods and/or new data sources and are therefore in line with the FSO's data innovation strategy and the Confederation's multi-annual programme for federal statistics. This site contains descriptions of the (pilot) projects currently being developed.

By publishing them we can involve users and partners at an early stage for both the development and consolidation of projects.

The aim of these statistical projects is to better meet users' needs in terms of efficiency, quality and speed. However, these statistics still have the potential to evolve, especially regarding their methodology, which is still being assessed. For this reason they are clearly marked as experimental and carry a logo that can easily be recognised.

Published statistics

Small area estimation (communes) of economic activity rate in the structural survey

The structural population survey provides important information on the population, including information about work. The whole purpose of Small Area Estimation is to push the boundaries imposed by standard methods.

The study showed that it is possible to obtain reliable estimates for both annual economic activity rates for communes that had a sample of at least 100 people.

Pilot projects within the data innovation strategy

On 21 November 2017, the FSO published its [data innovation strategy](#) (1).

This document is the FSO's first response to the wider subject of digitalisation. More specifically, it focuses on the application of complementary analysis methods (e.g. predictive analysis using advanced statistical techniques, data science and machine learning) that enable the current production of official statistics to be increased or completed. Five pilot projects have been chosen to implement this strategy and are in progress. Each project is described below.

Project 'Area Statistics Deep Learning' (ADELE)

The FSO's land use statistics are an invaluable tool for long-term land observation. This project involves learning and mastering the use of artificial intelligence (AI) technologies to eventually automate (even partially) the visual interpretation of aerial images in order to detect and classify changes.

Project 'Automation of NOGA coding' (NOGAUTO)

Automation of the coding of the economic activity of enterprises using Machine Learning methods applied to data already available within the FSO (data from surveys, descriptions in the commercial register, keywords, explanatory notes for classifications etc.) is supporting.

Sources:

Data Innovation Strategy 1.0 : <https://www.bfs.admin.ch/bfs/en/home/news/whats-new.gnpdetail.2017-0673.html>

Pilot Projects : <https://www.experimental.bfs.admin.ch/en/>

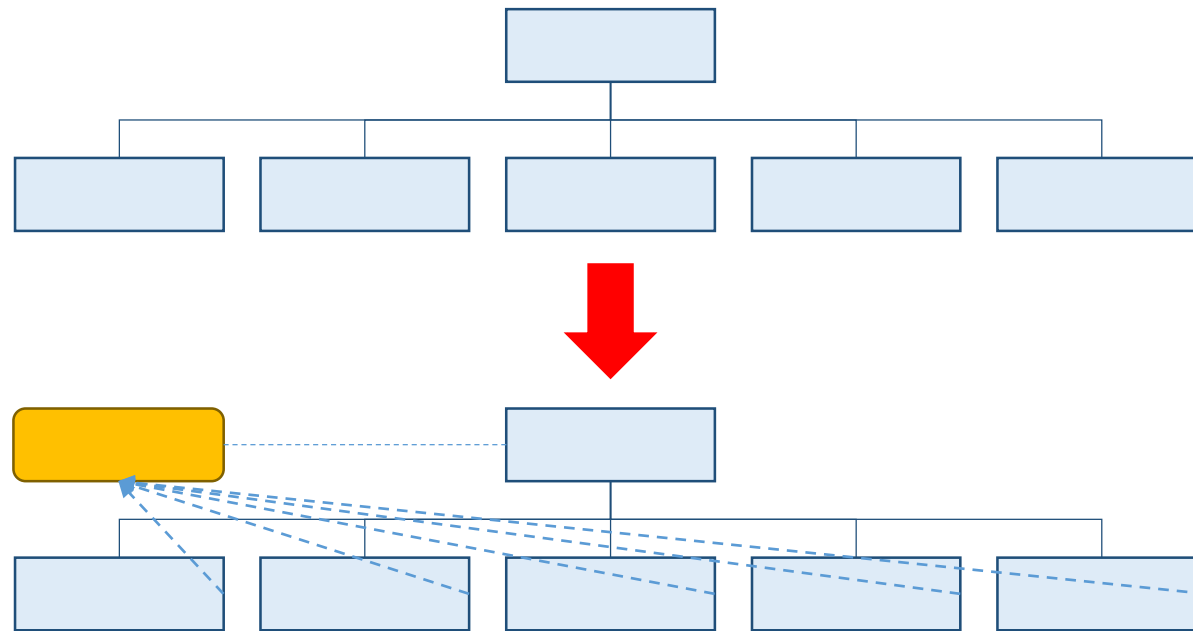


Agenda

- 1. What did we do ?**
2. What are our experiences ?
3. What will be the next major step ?



We have created an **ad-hoc** agile organization **inside** FSO - #1



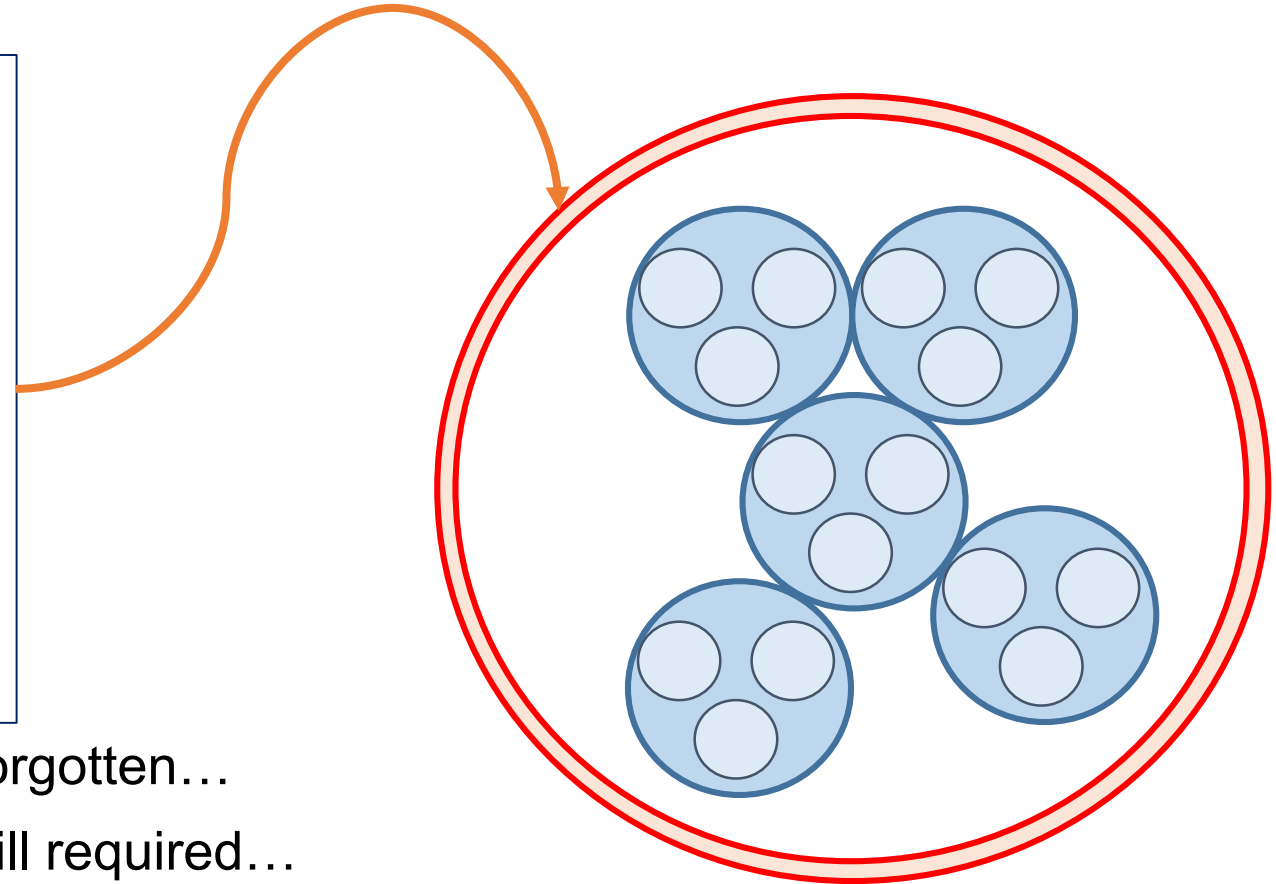
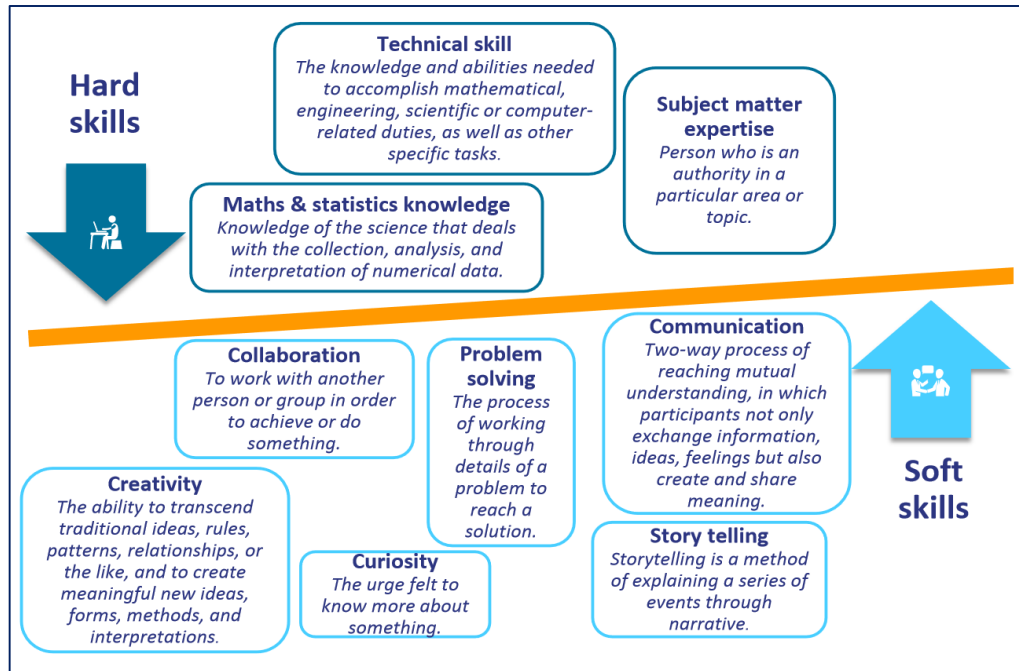
26 people (Statisticians, Methodologists, IT Specialists)
aged between 26 and 58 years old !

Rules

1. Ad-hoc organization is under the **final responsibility** of a board member.
2. Members of the ad-hoc organization have been **stayed subordinated** in their line organization !
3. **Five pilot projects** are managed by the ad-hoc organization. The final goal is to **go into production** at the end of 2019.



1. We have created an ad-hoc **agile** organization **inside** FSO - #2



- Classical hierarchy that needed to be forgotten...
- High degree of autonomy was and is still required...
- Choosing the right people at the beginning was not easy...



2. We have complemented their skills with **off-** and **on-the-job** teaching - #1

Off-the-job (9 x 1 day = **9 days** dispatched over 3 months)

- R for Data Science
- Python for Data Science
- Selected topics in data preparation/preprocessing
- Dimensionality reduction (e.g. PCA)
- Clustering (e.g. k-means)
- General introduction to supervised methods
- Regression (e.g. linear, logistic)
- Classification (e.g. k-NN, support vector machines)
- Decision trees
- Bayesian learning
- Neural networks
- Deep neural networks (e.g. CNN, RNN and LSTM)
- Model evaluation
- Feature selection



2. We have complemented their skills with off- and **on-the-job** teaching - #2

On-the-job (over 2 years)

Scientific advisors

- Professor for Data Science (University of Geneva) and the FSO's deputy head of methods have **accompanied/supervised** the five pilot projects and **validate** the results from the scientific point of view.

Team of “Data Scientists”

- Key idea is to **share/transfer** the skills and the knowledge **inside** the FSO Organization.
- Trying to avoid an organizational **split** between “traditional statisticians” and “modern statisticians or data scientists”.



Agenda

1. What did we do ?
- 2. What are our experiences ?**
3. What will be the next major step ?



What are our experiences ? - #1

Hard skills

- Math & statistical skill, Technical skill and subject matter can be learned. These skills are not always easy to acquire but it seems **not to be the hard part of the game!**
- Up to 2020 we will integrate a **standard curriculum “Data Scientist”** in our internal teaching program. We will collaborate more with Universities to define the exact content.

Soft Skills

- Soft skills (e.g. collaboration, creativity, problem solving, communication, story telling) need **a change in the mindset** of all employees.
- These skills are not so easy to teach... they primary need to be practiced on a daily basis
- Up to 2020 we will open a new mandatory curriculum **“Agile Organization”**.



What are our experiences ? - #2

Academic curricula

- Swiss Universities offer standards curricula in data science. The content of these curricula often does not match with the FSO's expectations.
- FSO does not really need “Supermen or Superwomen” that can cover all the production process on his own.

National Statistical Institute as employer

- Recruiting Data Scientist for a NSI likes FSO is not easy. Salaries are not so attractive and the tasks that we can offer to a young data scientist do not offer the expected variety and freedom that they are looking for.
- FSO wants to empower its employees and not to replace them with Data Scientist!



Agenda

1. What did we do ?
2. What are our experiences ?
- 3. What will be the next major step ?**



What will be the next major step ?

Version 2.0 of our Data Innovation Strategy

The FSO is currently defining a version 2.0 of its data innovation strategy for the years 2020 - 2023. The creation of an official **Data Innovation Lab** and a **Competence Center for Data Science** will (probably) be at the heart of the new strategy.

Work in progress !

Strategic objective 1: To create a central, cross-departmental and independent "**Data Innovation Lab**" (DIL) with a focus on **data innovation services**, *i.e.* the application of “complementary analytics methods” to data (and not to the type of data source and/or technology), for the FSO (I), the whole Swiss public statistics system (II) and the whole federal administration (III).



Questions & Answers

