"Show me your code, then I will trust your figures"

Towards software-agnostic open algorithms in statistical production

***Quality* Conference 2018**

J.Grazzini, P.Lamarche, J. Gaffuri & J.-M. Museux

Paradigm change for the production of Official Statistics

- new data source, combination of data: **data-centric approach**
- new algorithms /models and technologies: **more automation, metadata-driven & advanced analytics**
- privately owned data, IoT data: **remote computation & smart statistics**
- market competition *vs.* OS value added: **quality & transparency**
- new timely demands, data-informed decision-making: **agile data workflow & user-driven**

European Commission

outline: think global, code local…

- **Scope**: some banalities and many keywords
- **Walk the talk**: more talk and little walk
- **Thinking forward**: some discussion, few ideas and little action
- **Conclusion**: no solution, more questions

European Commission

This is not just "code"...

English EN

European Commission › Departments and executive agencies › Informatics ›

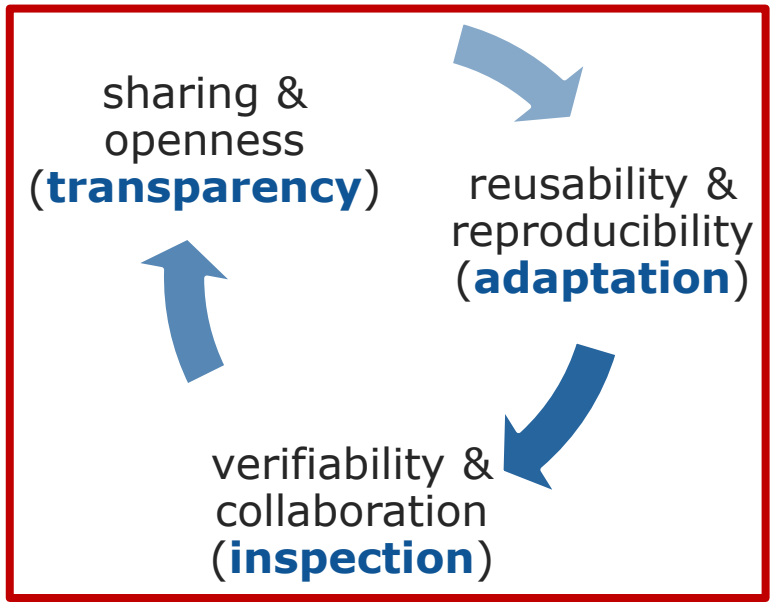
Open source software strategy

The European Commission will further increase the role of open source software for many of its key ICT services and software solutions. The renewed strategy puts a special emphasis on procurement, contribution to open source software projects and providing more open source software in the Commission.

the four drivers of digital social innovation are open data, open networks, open knowledge and open technology. The latter, in the form of software or hardware, allows digital solutions to social challenges to be shared more widely.

- but also **consistency & verifiability**
- ... **control & maintenance**
- ... **traceability & auditability**
- ... **accountability & reputation**

Open (data &) code and decision-making

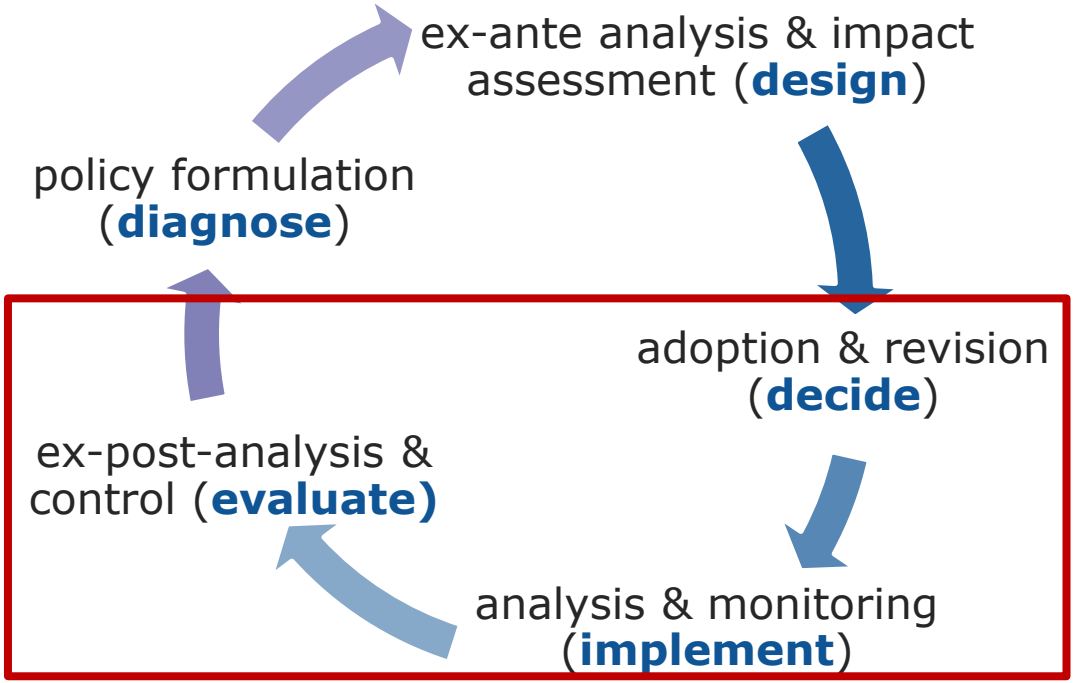


agile development

vs.

policymaking cycle

- **efficiency & timeliness**
- **transparency & collaboration**
- **quality & trust**



Q2018

European Commission

Open (& shared) code: *quid*?

- "**Open algorithm**" rather than "Open source software".

- "**Open source software**" are obviously preferred – though also susceptible to downside…

but **legacy proprietary software** are still in prominent use

- Best (consensual) practices from "**Open source community**":
  - **Openness**
  - **Reproducibility**
  - **Verifiability**
  - **Sharing**
  - **Reusability**
  - **Collaboration**

"What can I do you for?" Eurostat role to support open code (& software) (1/2)

Funding Agency Recommendations

from: V.Stodden, **"The reproducible research movement in statistics"**, 2013 (**https://web.stanford.edu/~vcs/talks/ISI-Aug302013-STODDEN.pdf**)

- Software and data should be "open by default" and access only restricted if openness conflicts with other considerations such as confidentiality.
- Add 'Reproducible Research' to the list of specific examples that proposals could include in their Broader Impact statements.
- Software and dataset curation should be explicitly included in grant proposals and recognized as a scientific contribution by funding agencies, and funds made available to support it.

European Commission

"What can I do you for?" Eurostat role to support open code (& software) (2/2)

Remark on deliverables and the use of the results of the action in relevant work packages

With reference to the aforementioned deliverables, the results of the action in the form of any kind, including software, should be made available to be fully exploited within the ESS. Therefore, the relevant source code or software applications that were developed as part of the action in fulfilling its objectives should be an implicit part of the reports for all aforementioned work packages respectively.

In fact, all deliverables should include the source code, testing material, and relevant documentation (requirement specification, design document, test plan and test log; administrative and user manuals) stored in a common public repository e.g. Github Eurostat domain.

Documentation should include the relevant information in a level of detail fitting to a prototype development serving the purposes of the action. Proofs-of-concept and prototypes should be possible to be deployed in the Infrastructure implemented by the project Big Data Test Infrastructure wherever applicable, provided that the infrastructure has been put in place and would offer the necessary functionalities. Wherever it would be applicable, part of the deliverables should be the assessment of the adequacy of the infrastructure with regard to its definition and design.

The use of mature and common free and open source technologies is encouraged, except when openness conflicts with other (recognised) priority considerations for software implementation.

EUROSTAT GRANTS 2018

in:

Call for proposals on
"*Multipurpose statistics and efficiency gains in production*"
(Call: ESTAT-PA11-2018-8)

The adoption and integration of such technologies is in line with both the Open Source Software Strategy of the European Commission[13] and the recommendations of the framework on Interoperability Solutions for Public Administrations, Businesses and Citizens[14]. It also supports the ESS Vision 2020[15] regarding the sharing of - software, but not only - resources between National Statistical Institutes.

Given this context, proposals making use of free and open source technologies will be considered in a good light. Submissions that explicitly promote the development of fully tested and documented source code and the sharing of reproducible and reusable software solutions will be evaluated positively.

Delivered software solutions aiming at integrating modular and scalable features and supporting different target architectures, so as to be shared within the ESS network, but also exposed to a larger community, will be preferred. Whenever possible, they should be made available also under the European Union Public License (EUPL)[16].

outline

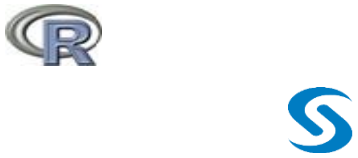
- **Objective**: some banalities and few keywords
- **Walk the talk**: more talk and little walk
- **Thinking forward**: some discussion, few ideas and little action
- **Conclusion**: no solution, more questions

European Commission

https://github.com/eurostat/**quantile**

- ✓ ***Agnostic:*** traditional quantile estimation technique is implemented **robustly on different platforms**.
- ✓ ***Controlled:*** parameters are not ad-hoc anymore but are reviewed to **correspond to state-of-the-art literature**.
- ✓ ***Serviced:*** web-app as a **plug & play quantile estimation service** so that users can focus on the estimation methods.

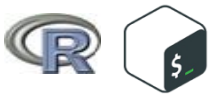
https://github.com/eurostat/**ICW**



- ✓ ***Reproducible and verifiable:*** the Experimental Statistics can be reproduced, producing the **same results from the same inputs**.
- ✓ ***Reusable:*** the code can be rerun and **used in new experiments**.

European Commission

https://github.com/eurostat/**PING**



- ✓ ***Proprietary software*** but ***open code.***
- ✓ ***Granular, modular, agnostic.***
- ✓ ***Versioned*** and ***documented:*** enhances **reproducibility**, enforces **quality assurance**.
- ✓ ***Tested*** and ***exemplified:*** supports **sharing and reuse** of modules, guarantees **reliability** and prepares future **migration**.

https://github.com/eurostat/**udoxy**

- ✓ ***Generic, agnostic:*** provide a framework to **document** stand-alone programs implemented in various programming languages.

European Commission

https://github.com/eurostat/**java4eurostat**

- ✓ ***data-centric:*** provides access to Eurostat data layers. Built on top of Eurostat **APIs and web-services**.
- ✓ ***Modular, generic,*** and ***reusable****:* **not application specific**, from low-level to advanced usage.
- ✓ ***Versioned*** and ***documented***.

https://github.com/eurostat/**Nuts2json**

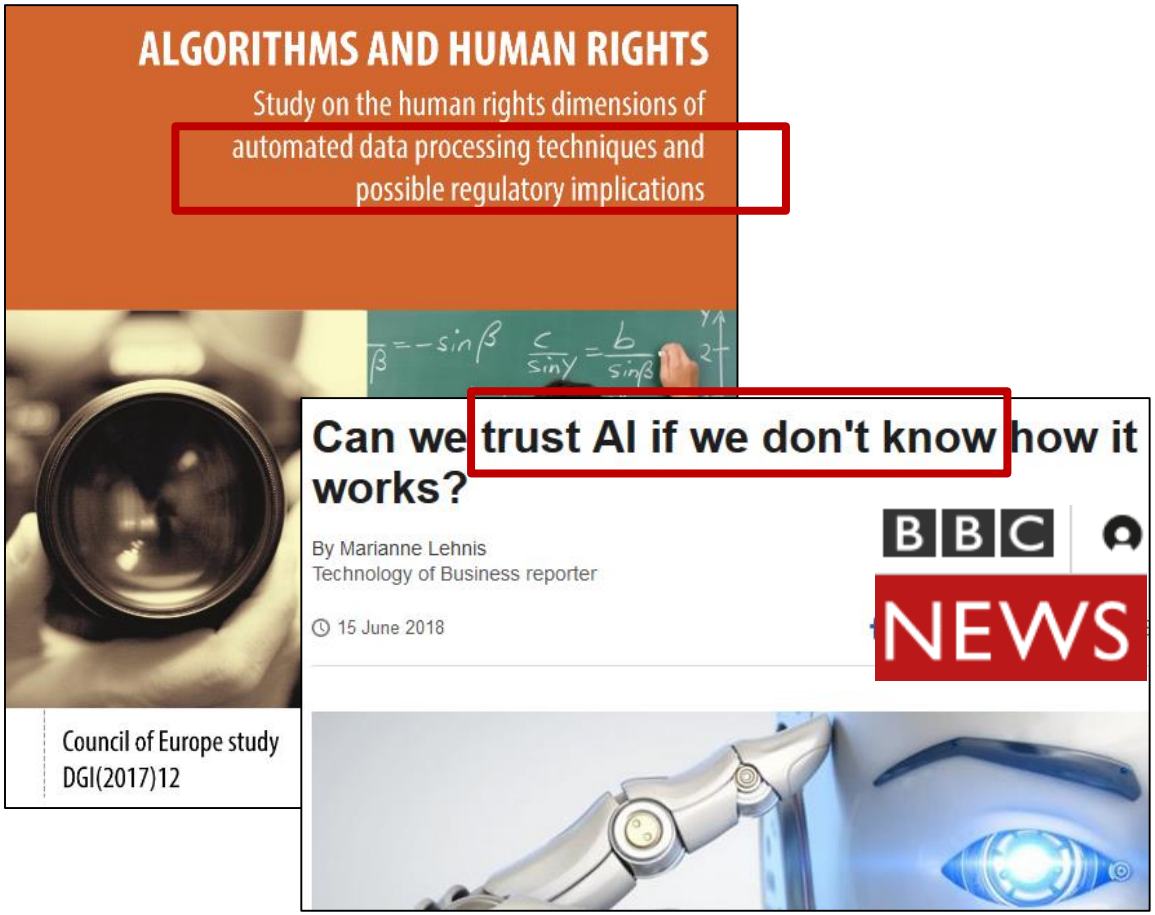
- ✓ ***data-centric:*** provides access to NUTS geometries for web mapping applications.
- ✓ ***Modular, generic,*** and ***reusable***.
- ✓ ***Versioned*** and ***documented***.

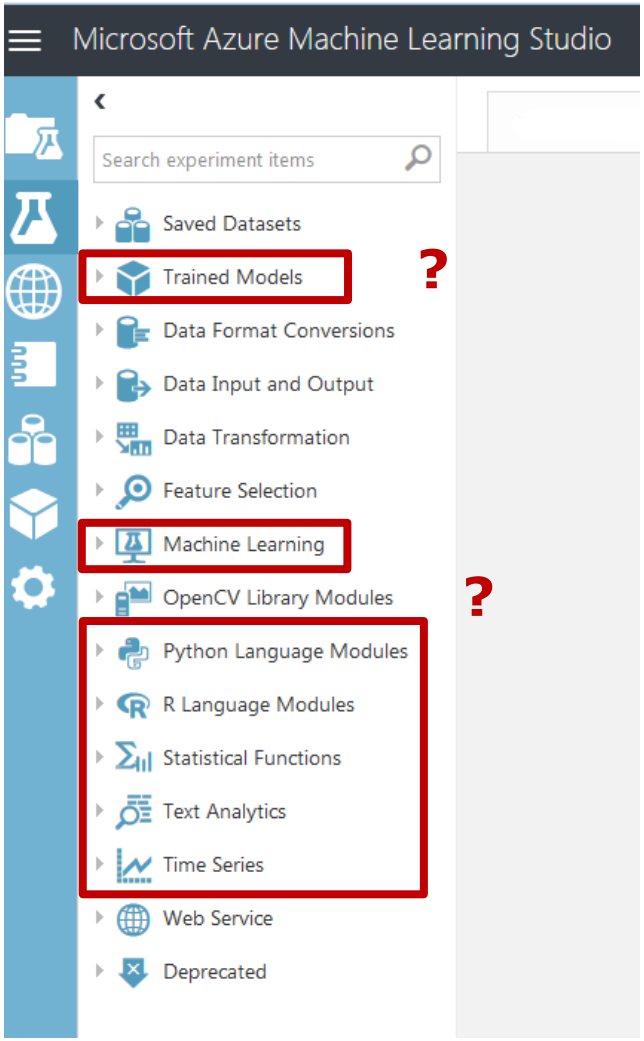
outline

- **Objective**: some banalities and few keywords
- **Walk the talk**: more talk and little walk
- **Thinking forward**: some discussion, few ideas and little action
- **Conclusion**: no solution, more questions

European Commission

Open data and open algorithms may not be enough





Open (& shared) statistical workflows: *quid*?

- Enable computational processes to be **run the exact same way in any environment**.

- Provide the computational components needed to **generate the same results from the same inputs**.

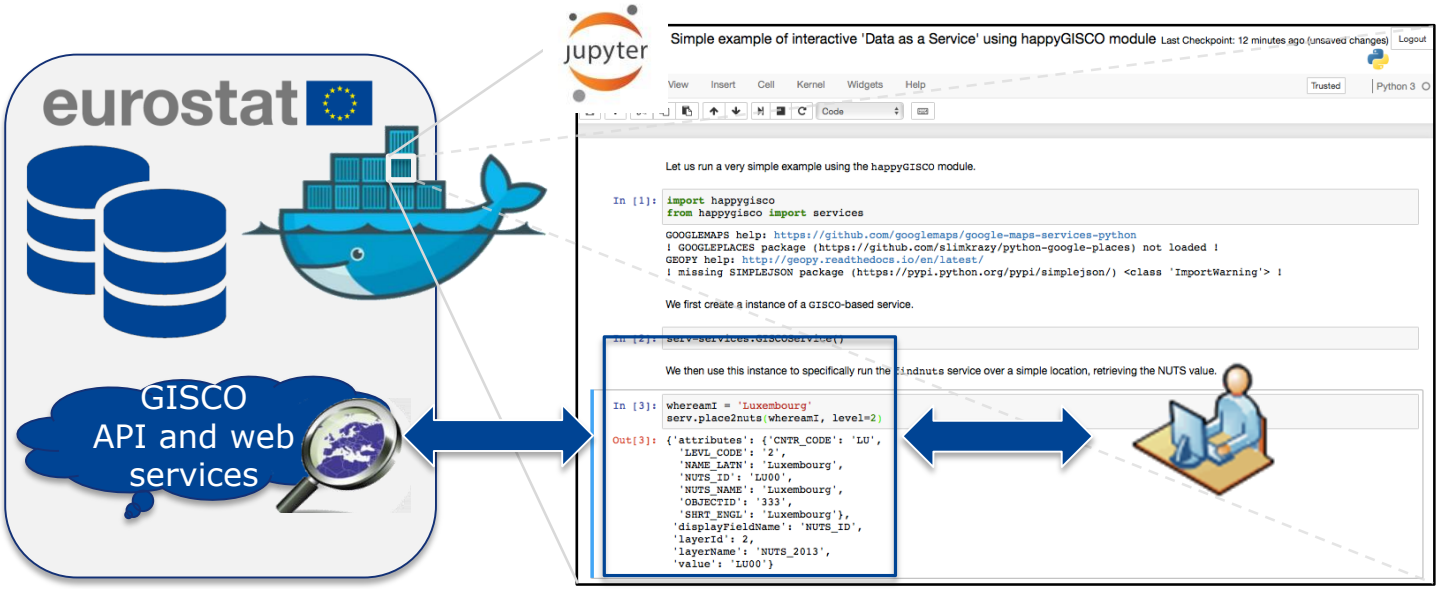
- Provide the public with further **insights into the workings of decision-making systems** to "judge for himself".

- **Participative** with incentives for **"produsers"** to **share back their analysis** for the benefit of the community.

European Commission

https://github.com/eurostat/**happyGISCO**

- ✓ ***Data-centric:*** Built ontop of Eurostat flexible **APIs and web-services**.
- ✓ ***User-driven:*** Provide versatile **interactive computing notebooks**.
- ✓ ***Agile:*** Distributed through lightweight platform independent **virtualised containers**.



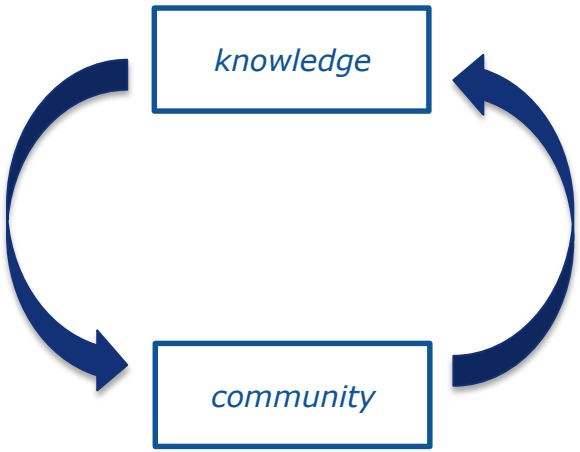
Q2018

outline

- **Objective**: some banalities and few keywords
- **Walk the talk:** more talk and little walk
- **Thinking forward**: some discussion, few ideas and little action
- **Conclusion**: no solution, more questions

Towards open data/algorithms/workflows...

- vision:
  - **Quality and trust** are fostered by **openness and transparency**.
  - Users/producers become "***produsers***".



- model:
  - **Open**, **shared**, and **collaborative**.
  - **Auditable, accountable** and **verifiable**.
  - **Agile**, **flexible**, and **continuous**.

- practice:
  - Today's technological solutions support an approach where **open algorithms and data are delivered as interactive, reusable and reproducible computing services**.

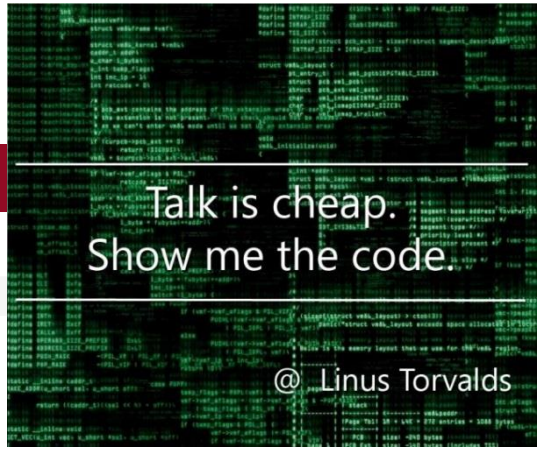
… and backwards same old (open) issues

- processes (development):
  - **Testing and certification** of statistical algorithms (sound methodology) and IT components (efficient implementation) ?
  - **Quality control and assessment** (actors: Eurostat, NSIs, larger community, …)?
  - **Maintenance of releases and versioning** (governance)?

- system (deployment):
  - **Integration of multiple data source and workflows**?
  - **Automation and transition** (migration) from research-grade experiments **to corporate production**?
  - **Audit trail**: reduce risk/cost of testing thanks to produsers?

European Commission

Thank you!



Communication of statistics in post-truth society: the good, the bad and the ugly

nature
International journal of science

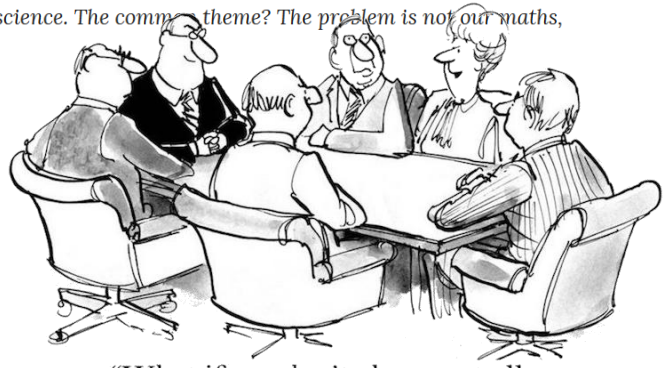
ent Research

n Research Analysis Careers Books & Culture

COMMENT · 28 NOVEMBER 2017

Five ways to fix statistics

As debate rumbles on about how and how much poor statistics is to blame for poor reproducibility, Nature asked influential statisticians to recommend one change to improve science. The comm[illegible] theme? The problem is not our maths, but ourselves.



"What if we don't change at all ... and something magical just happens?"

Publish your computer code: it is good enough

Freely provided working code — whatever its quality — improves programming and enables others to engage with your research, says **Nick Barnes**.

nature
International journal of science

Altmetric: 461 More detail »

Perspective | Published: 22 February 2012

The case for open computer programs

Q2018