Reproducible Analytical Pipelines & their value in Data Science

Ian Banda

Data Scientist

19 February 2025

Data Science Campus

Overview

- Pillars of Data Science in Official Statistics
- What are Analytical pipelines
- RAP principles
- Fundamental needs for Data Science in NSOs
- RAP in Data Science: a blueprint for skills
- RAP mentoring approach
- RAP strategy
- When to use RAP
- When RAP can be hard
- Guidance

Pillars of Data Science in Official Statistics.

1. Basics

Reproducible Analytical Pipelines (RAPs), moving from production to development, exploration and insight. Improving Quality Assurance.

2. Additional Insights

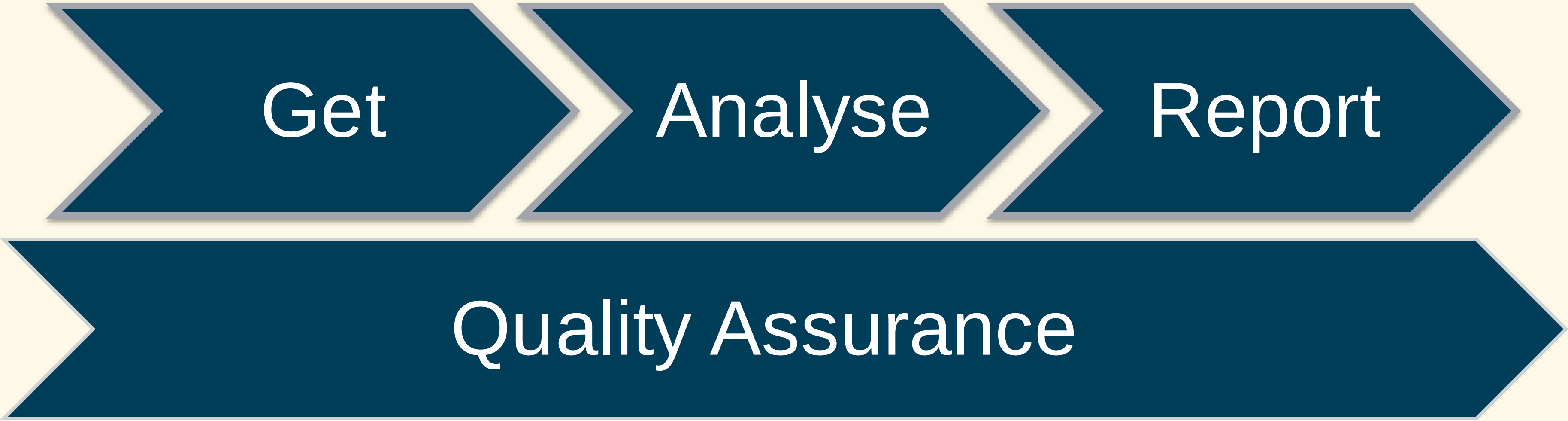
With capacity freed through increased automation ,can focus on enhancing Official Statistics through supplementary analysis that delivers deeper insights.

3. Transformation

New methods and data. Not necessarily ground breaking techniques, how can we create more useful statistics (eg. faster, more granular, higher quality) with modern data and tools.

Analytical pipelines

A process that produces an analytical product from data.



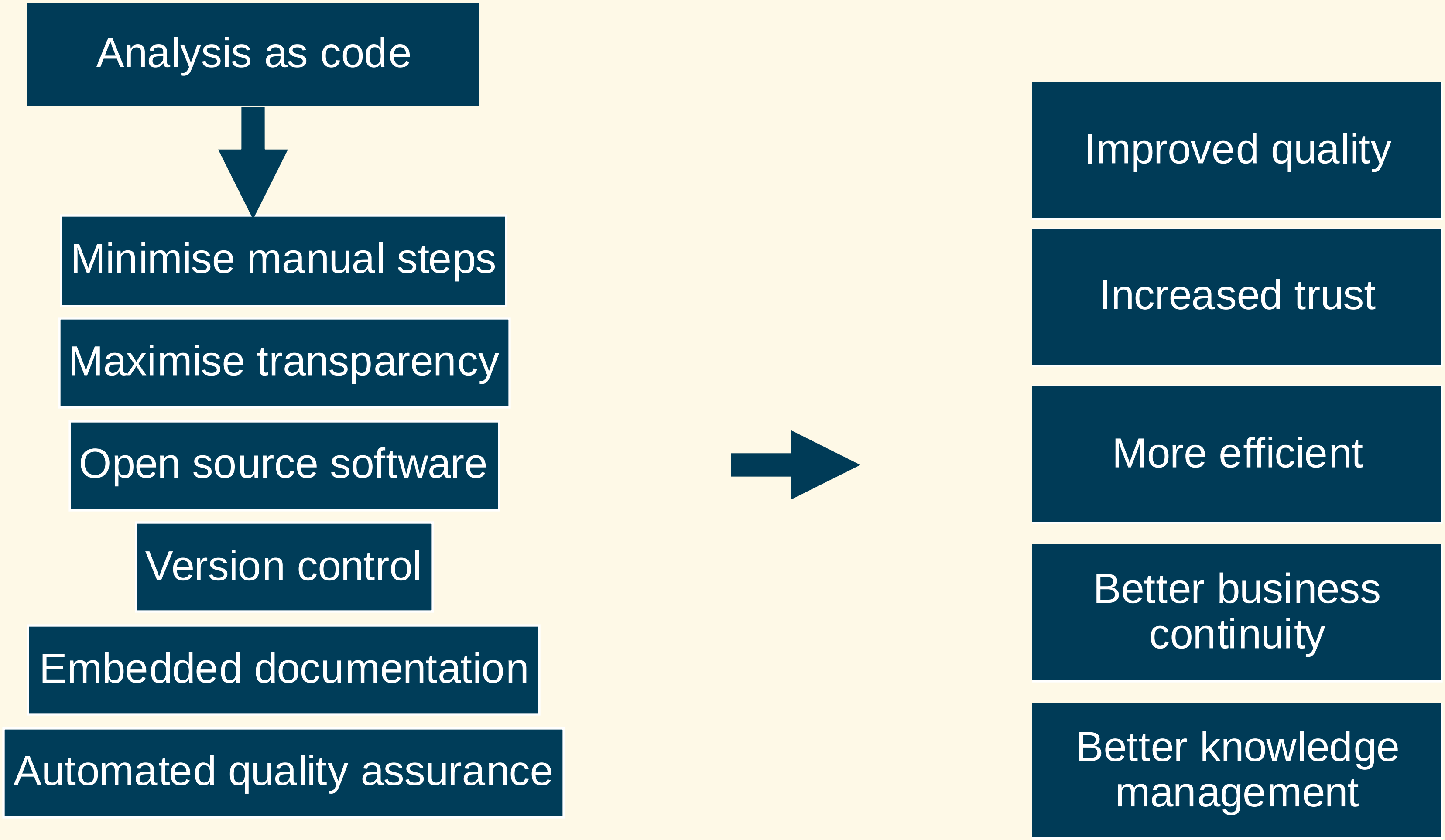
Get

Analyse

Report

Quality Assurance

RAP principles



Analysis as code

Minimise manual steps

Maximise transparency

Open source software

Version control

Embedded documentation

Automated quality assurance

Improved quality

Increased trust

More efficient

Better business continuity

Better knowledge management

Fundamental needs for Data Science in NSOs

1. **Skills**
   - Data literacy
   - Programming literacy
   - Following/building Good Practice

2. **Buy-in**
3. **Resource**

RAP in Data Science: a blueprint for skills

- Data Science ≠ RAP; RAP alone ≠ Data Science
- BUT; for RAP we need, e.g:
  - Understanding process/scope
  - Programming skills; R/Python
  - Focus on application & impact

RAP mentoring approach

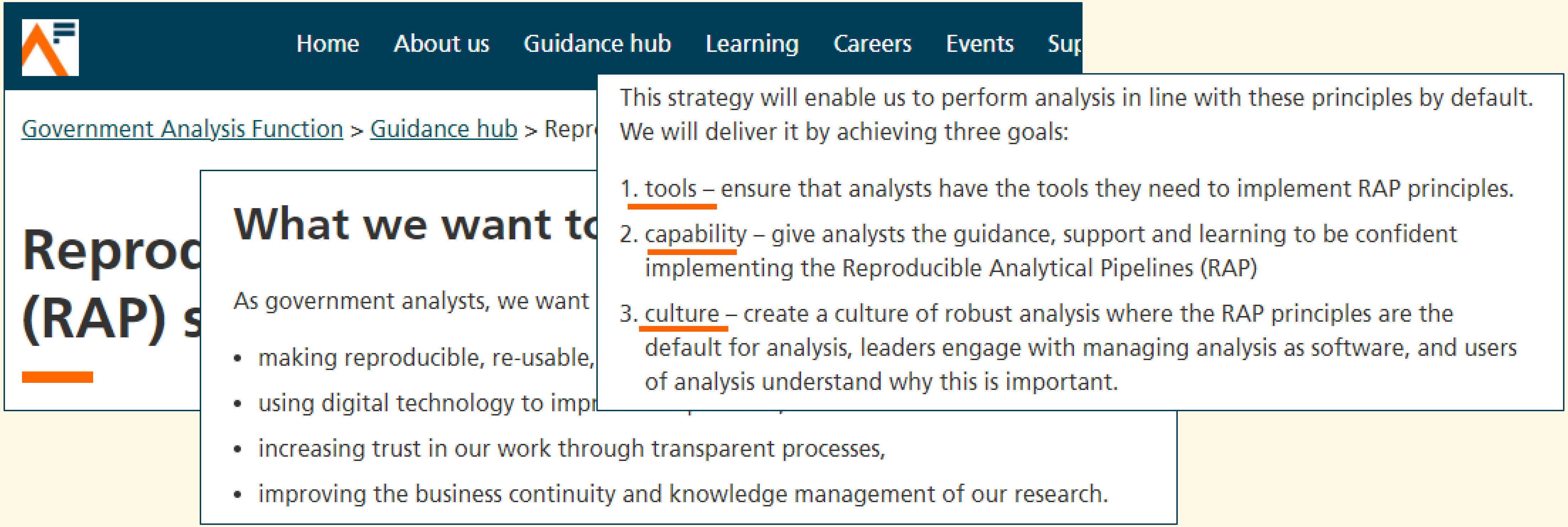
- Data scientist(s) mentoring small groups in partner NSO’s, e.g.
  - Scoping suitable work
  - Flexible & scalable training
  - Support pipeline development

- Longer period with regular check-ins

➔ Focus on ***application*** and ***impact***

RAP strategy

Standards provide framework for capability & Good Practice



Government Analysis Function > Guidance hub > Repr

Reproc
(RAP) s

What we want to

As government analysts, we want

- making reproducible, re-usable,
- using digital technology to impr
- increasing trust in our work through transparent processes,
- improving the business continuity and knowledge management of our research.

This strategy will enable us to perform analysis in line with these principles by default. We will deliver it by achieving three goals:

1. tools – ensure that analysts have the tools they need to implement RAP principles.
2. capability – give analysts the guidance, support and learning to be confident implementing the Reproducible Analytical Pipelines (RAP)
3. culture – create a culture of robust analysis where the RAP principles are the default for analysis, leaders engage with managing analysis as software, and users of analysis understand why this is important.

For this to work, we need…

- Commitment from senior managers
- Commitment from team members
- Enough time for team members to contribute
- A base level of technical understanding
- The right tools in the right place
- A plan to transition to business as usual

When to use RAP

- Your workflow is risky, time-consuming, hard to reproduce without manual intervention, or difficult to verify
- You want your analysis to be more efficient, more trustworthy and easier to quality assure
- It’s easiest to show value early on where data sources, processes and outputs are relatively stable
- The output doesn’t have to be a statistical report – it could be a standard set of statistics or graphs, a suite of data tables or a standard set of analyses and their outputs.

When RAP can be hard

- Limited access to open source tools. R, Rstudio, Rtools, Python, Git, GitHub
- Takes time to get to grips with techniques and tools – analysts need time to learn and deploy RAP
- When RAP is seen as “nice to have” rather than necessary by either managers and analysts
- When not enough time is set aside for RAP

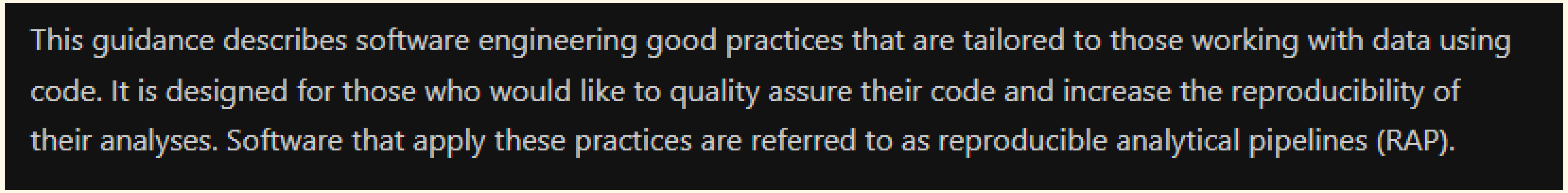
Summary, suggestions & discussion points

- RAP = efficiencies… but also blueprint for DS skills
  - Mentoring is efficient & scalable means to build both
- "Stepping stone" to Pillars 2 and 3

- Notes –
  - Some *initial* skills beneficial; e.g. precede with training courses?
  - Mentor & mentee ***availability is crucial*** (***buy-in***)
    - E.g. ring-fence part of staff time, but plan continued development
  - Take the long view: initial resource cost →→ efficiencies

Guidance

- ONS Data Science Campus
- UK Analysis Function RAP
- RAP Strategy (UK Analysis Function / ONS)
- RAP case studies
- Using RAP to improve statistics
- Quality Assurance of code for analysis and research



This guidance describes software engineering good practices that are tailored to those working with data using code. It is designed for those who would like to quality assure their code and increase the reproducibility of their analyses. Software that apply these practices are referred to as reproducible analytical pipelines (RAP).