Cabinet Office

Matthew Upson
Data Scientist
Government Digital Service
@m_a_upson

What is GDS?

GDS is here to make it easy for government to be digital

https://goo.gl/fLKXkS

*The other things data scientists need to know

* Some of...

All   Videos   Images   News   Shopping   More   Settings   Tools

About 39,900,000 results (0.65 seconds)

R vs Python for Data Science: The Winner is … - KDnuggets
www.kdnuggets.com/2015/05/r-vs-python-data-science.html ▼
In the battle of "best" data science tools, python and R both have their pros and cons. Selecting one over the other will depend on the use-cases, the cost of ...

R vs Python for Data Science: Summary of Modern Advances
https://elitedatascience.com/r-vs-python-for-data-science ▼
11 Feb 2017 - Should you choose R or Python for Data Science? Thanks to recent advances made in both languages, you really can't go wrong with either.

Choosing R or Python for data analysis? An infographic (Article)
https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis ▼
12 May 2015 - or "I need a fast solution, should I go for Python or R?". In today's post, I present to you our new Infographic "Data Science Wars: R vs Python", ...

#DataOps

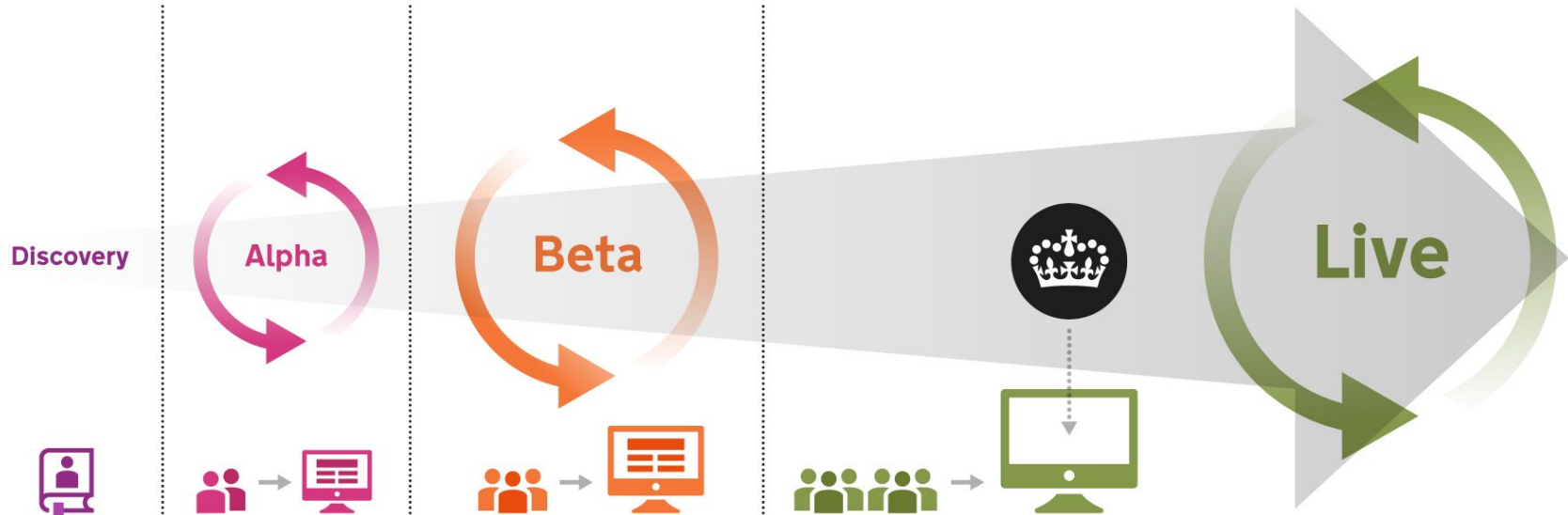dataopsmanifesto.org

Data science is a team sport

Data scientists often work in mixed agile[a] teams[b] to deliver projects[c]

GDS

Data scientists often work in mixed agile[a] teams[b] to deliver projects[c]

a. Knowing about agile is helpful

agilemanifesto.org

Data scientists often work in mixed agile[a] teams[b] to deliver projects[c]

b. You will need to work alongside other job roles (and speak a common language)

My team:
6x Data Scientist
2x Software Developer
1x Content Designer
1x User Researcher
2x Delivery Manager
1x Product Manager
2x Policy Analysts

Very technical

Less technical

Data scientists often work in mixed agile[a] teams[b] to deliver projects[c]

c. Projects are often things (services, applications, digital products), not just ad-hoc analysis or models.

Getting buy-in and scoping: c. 40%

Data science: c. 20%

Handover and engagement: c. 40%

GDS

Analytics is code

Make it reproducible

an example from government

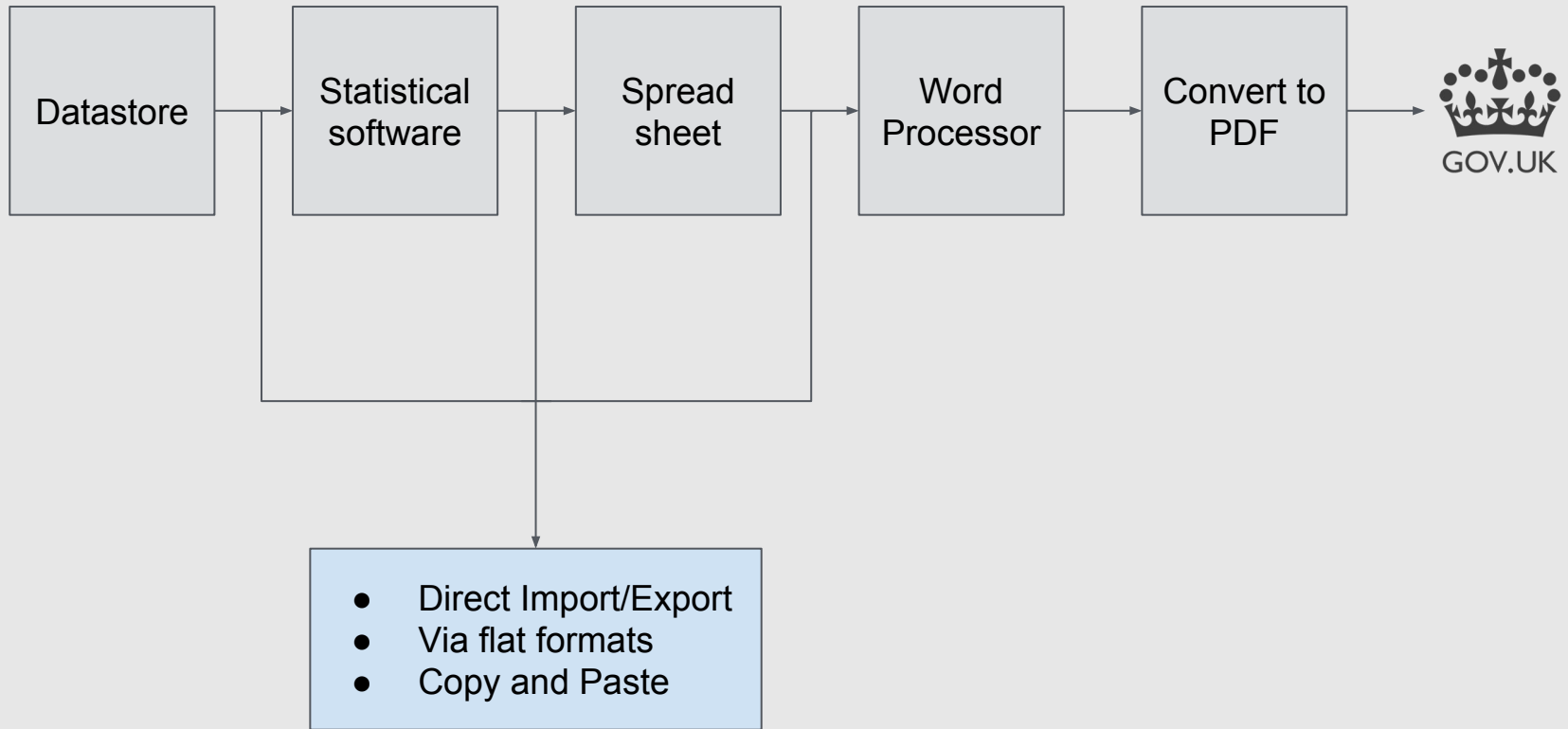The final **thing** is not the thing.

The code is the thing.

Code is easy to reproduce, **things** aren't.
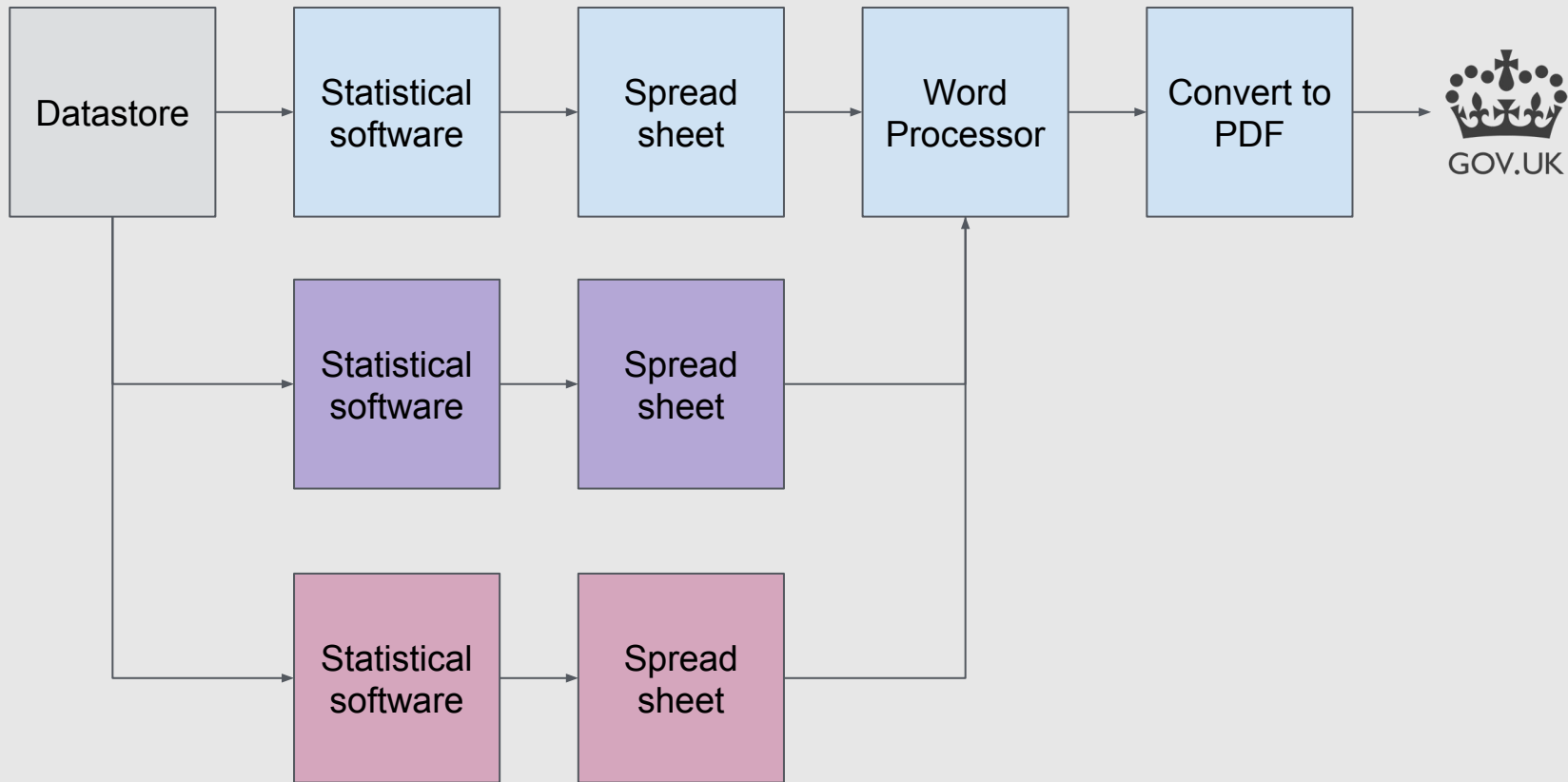
Home

Official Statistics

DCMS Sectors Economic Estimates 2016

From: Department for Digital, Culture, Media & Sport
Part of: DCMS Sectors Economic Estimates, Tourism, Arts and culture,
and Communications and telecomms
Published: 4 August 2016
Last updated: 1 November 2016, see all updates

Datastore → Statistical software → Spread sheet → Word Processor → Convert to PDF → GOV.UK

- Direct Import/Export
- Via flat formats
- Copy and Paste
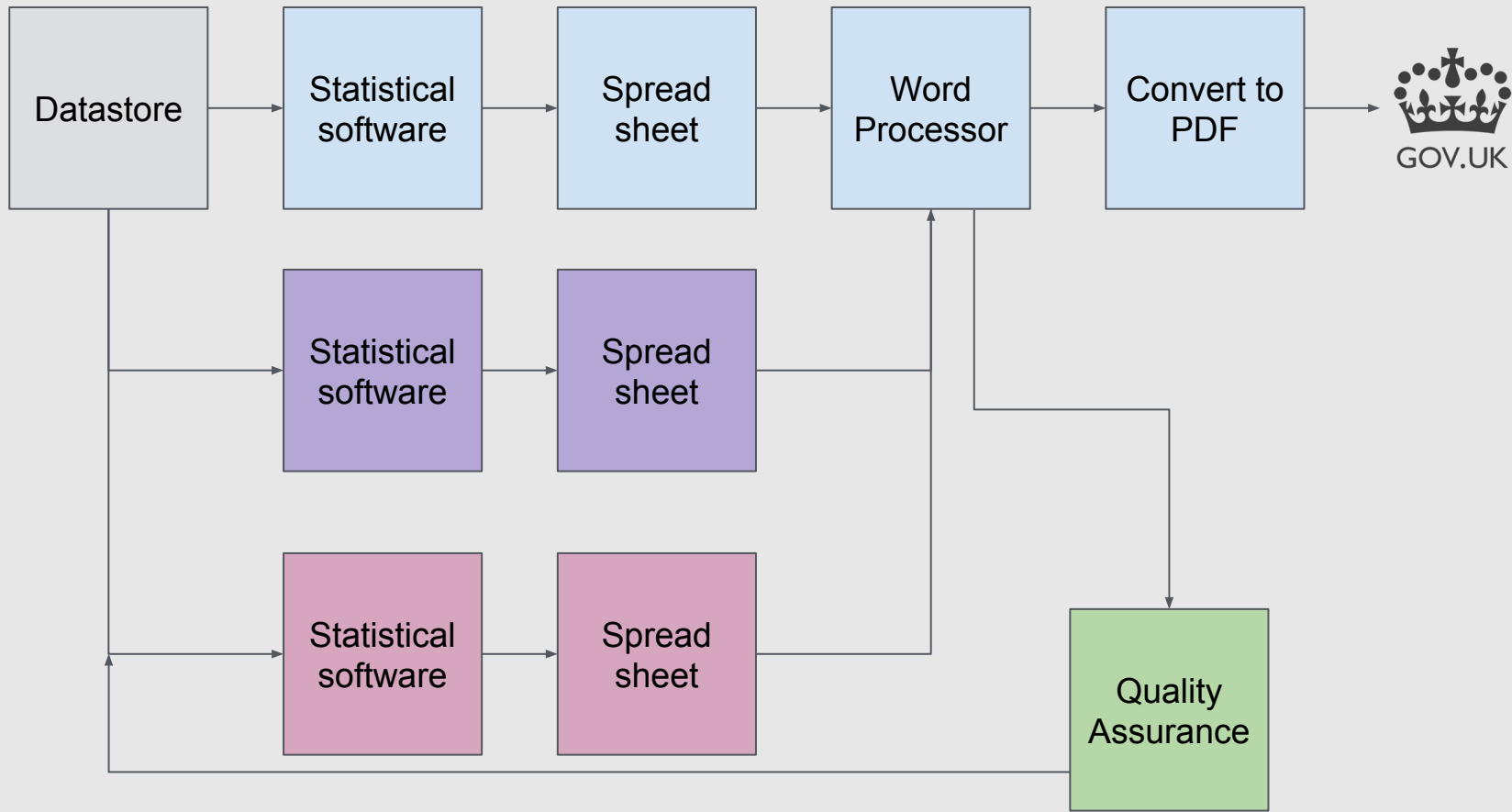
Applying DataOps principles reduces production time by c. 75% and improves auditability and quality.

1. Remove all manual steps from data manipulation

2. Version control everything

pharmacy_analysis_version_3.R
pharmacy_analysis_version_3a.R
pharmacy_analysis_version_3a1.R
pharmacy_analysis_version_3a1_MU_checked.R

This is not version control

pharmacy_analysis_version_final_checked_MU1.R
pharmacy_analysis_version_final_final_checked_MU1.R
pharmacy_analysis_version_really_final_this_time.R

Commits on Aug 8, 2017

Merge pull request #112 from ukgovdatascience/model-comments ...

ivyleavedtoadflax committed on GitHub 29 days ago ✓ 60.71%

2c68cf9 ‹›

Commits on Aug 2, 2017

Merge branch 'master' into model-comments

davidread committed on GitHub on Aug 2 ✓ 60.71%

cf78be4 ‹›

Added comments to model

davidread committed on Aug 2 ✓ 60.71%

962eba4 ‹›

Commits on Aug 1, 2017

Merge pull request #111 from ukgovdatascience/feature/majorityvote ...

davidread committed on GitHub on Aug 1 ✓

4ea424d ‹›

Commits on Jul 26, 2017

Update Urls model to fit database changes

ivyleavedtoadflax committed on Jul 26 ✓ 60.71%

Commits on Jul 23, 2017

Update README.md

ivyleavedtoadflax committed on GitHub on Jul 23 ✗ 60.58%

3. Write software, not scripts

http://r4ds.had.co.nz/functions.html

Put it into a package

eesectors

- Functions
- Tests
- Documentation
- Data
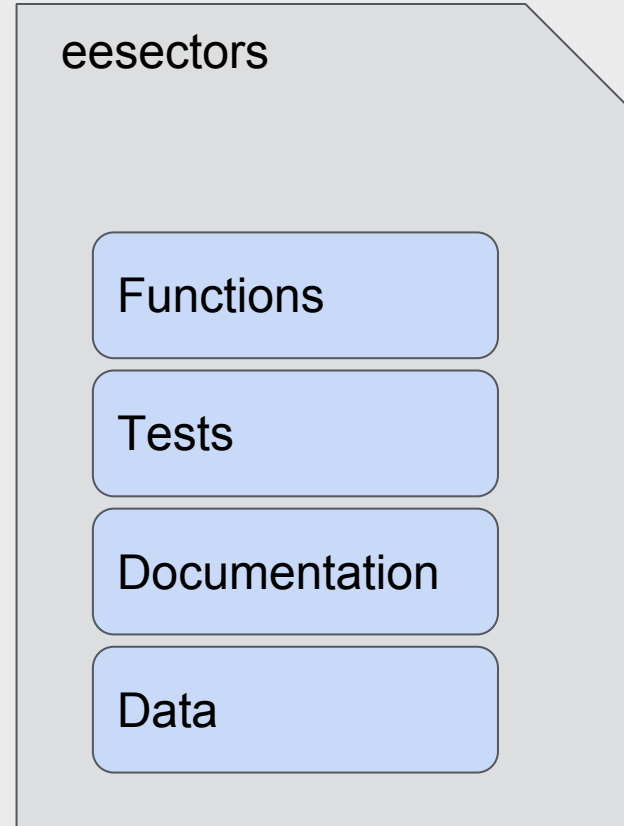
4. Test the critical things over and over again (and automate it)

Data integrity

Does the data look right?

Unit tests

Does this bit of code do what I expect?

testthat for R
pytest for Python

End-to-end tests

Does the whole process work as expected?

Automate it!

✓ **Pull Request #91**   Fix typo in README

○─ Commit 8a5e1ec ⬈

⑂ #91: Fix typo in README ⬈

⑂ Branch master ⬈

Ⓞ Matthew Upson authored and committed

⑂ **#703 passed**

⏱ Ran for 16 min 46 sec

📅 21 days ago

↻ Restart build

5. Test to see what you have tested
(and automate that too)

'Code coverage' percentages

📄 extract_GVA_data.R	100.0...
📄 extract_SIC91_data.R	100.0...
📄 extract_tourism_data.R	100.0...
📄 figure3.1.R	92.85%
📄 figure3.2.R	93.18%
📄 figure3.3.R	92.85%
📄 maha_check.R	100.0...
📄 onattach.R	100.0...
📄 raise_issue.R	100.0...

7. Combine the narrative with the code (and version that too)

6. Share what you have learnt

Inbox — ma... | digital.cab... | RAP_RSS_... | RAP_pres... | RSS Data S... | RSS 2017 | RSS 2017 | RSS_2017... | RAP_semin... | feedback | ukgovdata...

Matthew

GitHub, Inc. [US] | https://github.com/ukgovdatascience/eesectors

Apps | GDS Useful Links | awslabs/deeplearni...

Other Bookmarks

This repository | Search

Pull requests | Issues | Marketplace | Explore

ukgovdatascience / eesectors

Unwatch ▾ 6 | ★ Star 21 | Fork 2

Code | Issues 10 | Pull requests 4 | Projects 1 | Wiki | Settings | Insights ▾

⚠ Prototype R package for the creation of the DCMS Sectors Economic Estimates statistical release

Edit

reproducible | government | statistics | dataops | Manage topics

155 commits | 5 branches | 13 releases | 2 contributors

Branch: master ▾ | New pull request | Create new file | Upload files | Find file | Clone or download ▾

ivyleavedtoadflax committed on GitHub Merge pull request #85 from ukgovdatascience/fix/extract_DCMS_sectors ··· Latest commit 5e91f35 on Mar 7

R	fixed extract_DCMS_sectors (closed #60)	6 months ago
README_files/figure-markdown_s...	Update README and DESCRIPTION	7 months ago
data	Add extract_DCMS_sectors and DCMS_sectors data	7 months ago
inst	Fix UK_GVA total intermediate issue	7 months ago
man	Re-factored part of appendSectors.R function	6 months ago

I am not original: **partiallyderivative.com**

Read the blog: **https://goo.gl/vYoEGD**

Matthew Upson
@m_a_upson