

Armchair Auditing: What could possibly go wrong?

Charlie Rahal

LCDS and Nuffield College, University of Oxford

Government Data Science Meetup



Upholding the Civil Service Code, lets remain impartial/apolitical.

In the spirit of 'Open Data', everything with free, and 'Open' tools:

- Linux (operating system)
- Python (data wrangling and statistical software)
- L^AT_EX (typesetting)

All the research you see here today is at: github.com/crahal.

Slides (crahal.github.io/files/GovDataScienceMeet.pdf):



How it started...

- UK always ~top of rankings re: govt. transparency.
- Coalition Govt. opened up procurement (2010).
- A totally new type of (semi)-structured data.
- From LAs to NHS hierarchy to central government.
- ‘Open Data’ is worth ~2€bn (European Commission)
- **Armchair Auditors** hold public sector to account!
- Enormous potential for **academic research...**

WHAT COULD POSSIBLY GO WRONG?

Lets list a few of the problems with procurement data.

Lets list a few of the problems with procurement data.

1. Some organizations don't believe in their (legal?) obligation.
(Some organizations under-staffed / have staff turnover)

Lets list a few of the problems with procurement data.

1. Some organizations don't believe in their (legal?) obligation.
(Some organizations under-staffed / have staff turnover)
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations across the internet)

Lets list a few of the problems with procurement data.

1. Some organizations don't believe in their (legal?) obligation.
(Some organizations under-staffed / have staff turnover)
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations across the internet)
3. As far as I can tell, this is **manually** entered at some point.
(Resulting in a great deal of typing mistakes in raw data: eek!)

Lets list a few of the problems with procurement data.

1. Some organizations don't believe in their (legal?) obligation.
(Some organizations under-staffed / have staff turnover)
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations across the internet)
3. As far as I can tell, this is **manually** entered at some point.
(Resulting in a great deal of typing mistakes in raw data: eek!)
4. Data commonly converted from a tabular file...

Lets list a few of the problems with procurement data.

1. Some organizations don't believe in their (legal?) obligation.
(Some organizations under-staffed / have staff turnover)
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations across the internet)
3. As far as I can tell, this is **manually** entered at some point.
(Resulting in a great deal of typing mistakes in raw data: eek!)
4. Data commonly converted from a tabular file... **into** a .pdf!
(This is a **direct** downgrade in the Berners-Lee 5* schema)

Lets list a few of the problems with procurement data.

1. Some organizations don't believe in their (legal?) obligation.
(Some organizations under-staffed / have staff turnover)
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations across the internet)
3. As far as I can tell, this is **manually** entered at some point.
(Resulting in a great deal of typing mistakes in raw data: eek!)
4. Data commonly converted from a tabular file... **into** a .pdf!
(This is a **direct** downgrade in the Berners-Lee 5* schema)
5. Naturally, of course, there are no supplier UIDs.
(That would make life too easy)

Lets list a few of the problems with procurement data.

1. Some organizations don't believe in their (legal?) obligation.
(Some organizations under-staffed / have staff turnover)
 2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations across the internet)
 3. As far as I can tell, this is **manually** entered at some point.
(Resulting in a great deal of typing mistakes in raw data: eek!)
 4. Data commonly converted from a tabular file... **into** a .pdf!
(This is a **direct** downgrade in the Berners-Lee 5* schema)
 5. Naturally, of course, there are no supplier UIDs.
(That would make life too easy)
- ⋮

Lets list a few of the problems with procurement data.

1. Some organizations don't believe in their (legal?) obligation.
(Some organizations under-staffed / have staff turnover)
 2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations across the internet)
 3. As far as I can tell, this is **manually** entered at some point.
(Resulting in a great deal of typing mistakes in raw data: eek!)
 4. Data commonly converted from a tabular file... **into** a .pdf!
(This is a **direct** downgrade in the Berners-Lee 5* schema)
 5. Naturally, of course, there are no supplier UIDs.
(That would make life too easy)
- ⋮

This could become a very long list.

DATA SCIENCE TO THE RESCUE!

DATA SCIENCE TO THE RESCUE!

Lets see an example of the **cleanest** procurement file I could find

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

Expenditure Over Threshold

7 / 17

Expenditure Over Threshold

[illegible]

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

7 / 17

What does a typical ETL pipeline look like in this domain?

What does a typical ETL pipeline look like in this domain?

1. Set up (thousands) of webscrapers.

What does a typical ETL pipeline look like in this domain?

1. Set up (thousands) of webscrapers.
2. Set up a parsing module to ingest heterogenous files.

What does a typical ETL pipeline look like in this domain?

1. Set up (thousands) of webscrapers.
2. Set up a parsing module to ingest heterogenous files.
3. Harmonize field names.

What does a typical ETL pipeline look like in this domain?

1. Set up (thousands) of webscrapers.
2. Set up a parsing module to ingest heterogenous files.
3. Harmonize field names.
4. Clean for typing errors, parsing mishaps, etc.

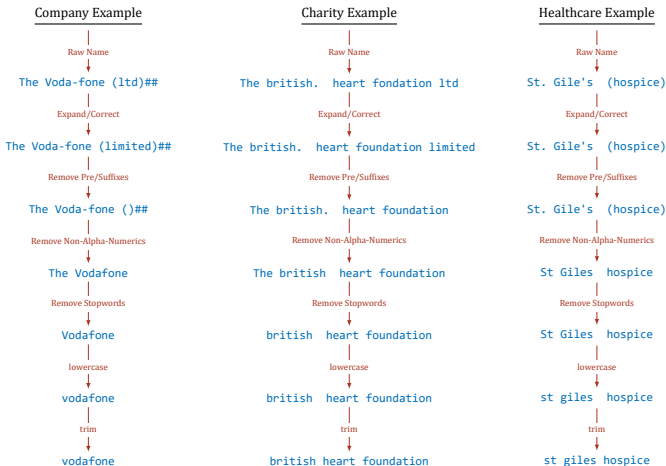
What does a typical ETL pipeline look like in this domain?

1. Set up (thousands) of webscrapers.
2. Set up a parsing module to ingest heterogenous files.
3. Harmonize field names.
4. Clean for typing errors, parsing mishaps, etc.
5. Reconcile against institutional registers.

What does a typical ETL pipeline look like in this domain?

1. Set up (thousands) of webscrapers.
2. Set up a parsing module to ingest heterogenous files.
3. Harmonize field names.
4. Clean for typing errors, parsing mishaps, etc.
5. Reconcile against institutional registers.
6. Analysis. Hurrah!

A key part to 'reconciliation' is 'normalisation':



This greatly helps increase the frequency and quality of matches.

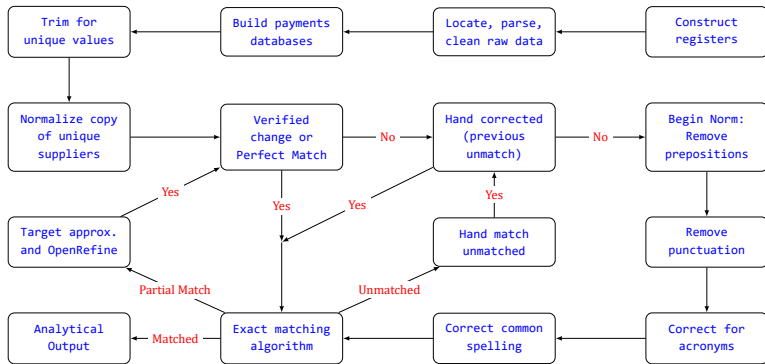
Example Reconciliation Issue

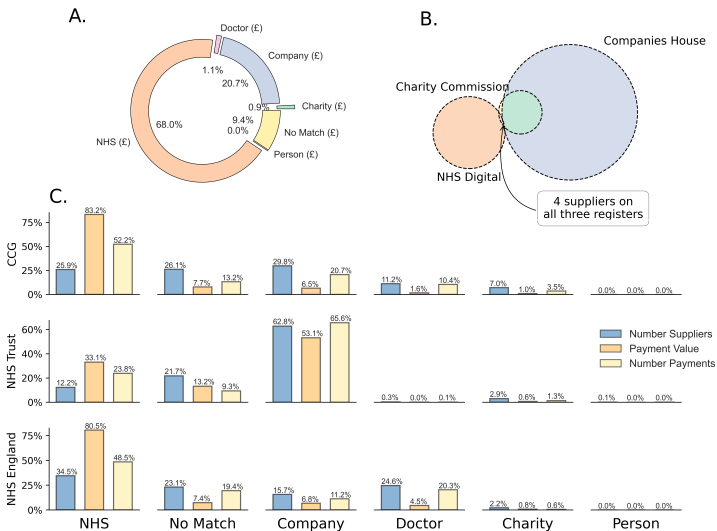
These are two real companies:

1. THE HEARTH CENTRE (HORSLEY) LTD
(Company number 04929060)
2. THE HEARTH CENTRE
(Company number 07219435)

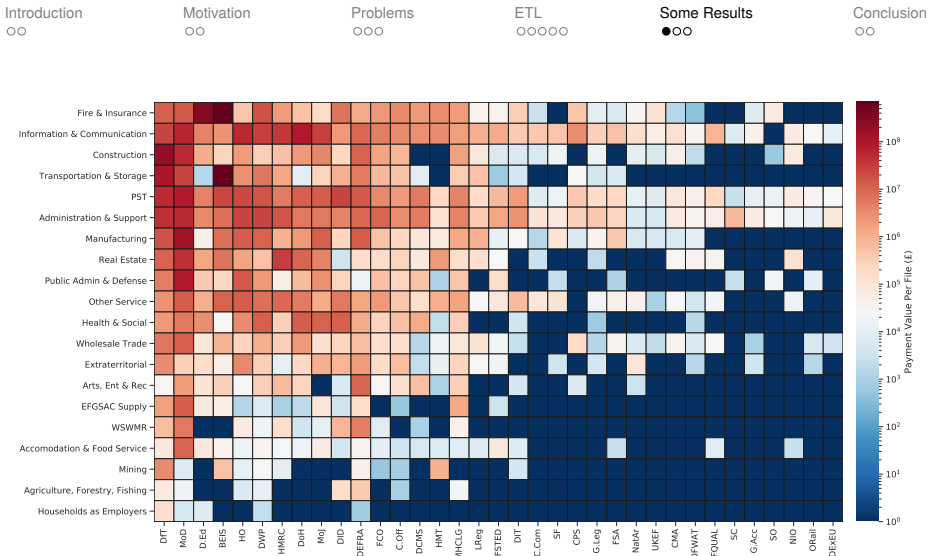
What happens if we normalize names as per the previous slide?

Many Possible Paths to Take

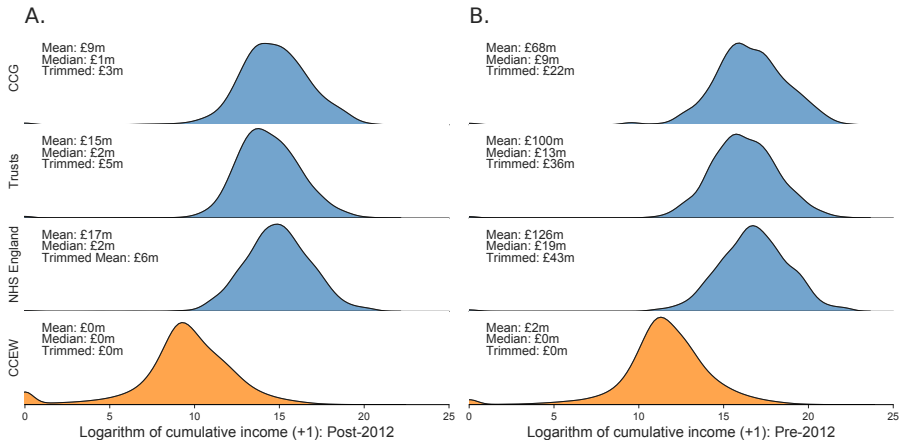




‘What can we match?’



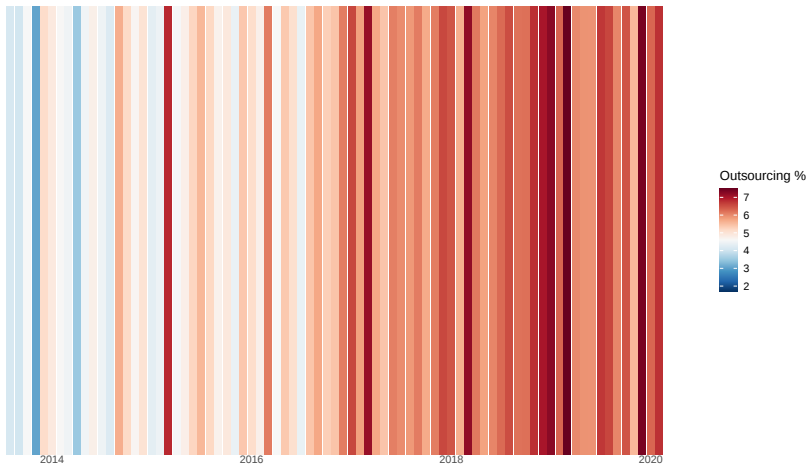
‘Tools for Transparency in Central Government Spending’



‘The Role of Non-Profits in Public Health Service Provision’

With John Mohan

Creeping Privatisation in England's NHS



Source: Raw CCG payments approximately matched to Companies House and the Charity Commission, then manually verified

‘Outsourcing health-care services to the private sector’

By Ben Goodair and Aaron Reeves

Conclusion

- This data is messy, and difficult to work with.

Conclusion

- This data is messy, and difficult to work with.
- Tools of data science provide an enormous help.

Conclusion

- This data is messy, and difficult to work with.
- Tools of data science provide an enormous help.
- Manual wrangling/reconciling at times unavoidable.

Conclusion

- This data is messy, and difficult to work with.
- Tools of data science provide an enormous help.
- Manual wrangling/reconciling at times unavoidable.
- The benefits of cleaning this data are (**still**) huge.

Conclusion

- This data is messy, and difficult to work with.
- Tools of data science provide an enormous help.
- Manual wrangling/reconciling at times unavoidable.
- The benefits of cleaning this data are (**still**) huge.
- **A lot** more to be done in this area!

Conclusion

- This data is messy, and difficult to work with.
- Tools of data science provide an enormous help.
- Manual wrangling/reconciling at times unavoidable.
- The benefits of cleaning this data are (**still**) huge.
- **A lot** more to be done in this area!
- We're currently finishing a dashboard.
 - Please get in contact if you'd like to be involved!

Thank you for attending!

Thank you to my collaborators!

Don't hesitate to get in touch at:

charles.rahall@demography.ox.ac.uk

and

github.com/crahall!