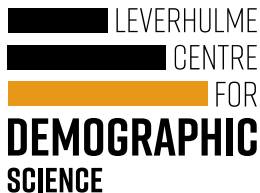


Armchair Auditing: What could possibly go wrong?

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)
- $\text{\LaTeX}$ (typesetting)

All the research you see here today is at:

github.com/crahal

and

crahal.github.io/files/Gov_Data_Science_Meetup.pdf

How it started...

- UK always ~top of rankings re: govt. transparency.
- David Cameron 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

7 / 17

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

Expenditure Over Threshold

Department of Health	NH	Beasey	CGG	10/04/2021	Clinical/Medical-Other Public Sector	COMMUNITY SERVICES	BESELY COUNCIL	11595633	658,002.00	6100071658
Department of Health	NH	Beasey	CGG	10/04/2021	Clinical/Medical-Other Public Sector	COMMUNITY SERVICES	BESELY COUNCIL	11595662	100,000.00	6100071661
Department of Health	NH	Beasey	CGG	10/04/2021	Other Gen. Supplies 1-5 yr	COMMUNITY SERVICES	BESELY COUNCIL	11595662	60,556.07	6100071661
Department of Health	NH	Beasey	CGG	10/04/2021	Clinical/Medical-Other Public Sector	Winter Resilience	BESELY COUNCIL	11595662	64,002.00	6100071687
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	CHC - COMMUNITY SERVICES	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11007201	65,258.00	610008684
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	CHC - COMMUNITY SERVICES	CRAFTFORD AND GRAVESEND NHS TRUST	11562030	65,258.00	610008684
Department of Health	NH	Beasey	CGG	10/04/2021	Clinical/Medical-Voluntary Sector	COMMISINGING - NON ACUTE	BESELY COUNCIL	11615604	30,000.00	6100070989
Department of Health	NH	Beasey	CGG	10/04/2021	Clinical/Medical-Independent Sector	ACUTE COMMISINGING	NHS HEALTH LTD	11615710	38,248.24	610026662
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	ACUTE COMMISINGING	NEWSTAY NHS FOUNDATION TRUST	11623784	38,248.24	610026662
Department of Health	NH	Beasey	CGG	10/04/2021	Clinical/Medical-Voluntary Sector	HOSPICES	GREENWOOD & BESELY COMMUNITY HOSPICE	11618137	218,000.00	6100071697
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	ACUTE COMMISINGING	GREENWOOD & BESELY COMMUNITY HOSPICE	11618137	218,000.00	6100071697
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PROGRAMME PROJECTS	NHS PROPERTY SERVICES LTD	11614167	218,714.71	6100070294
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PROGRAMME PROJECTS	NHS PROPERTY SERVICES LTD	11642664	25,675.00	6100070294
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	ACUTE COMMISINGING	CRAFTFORD AND GRAVESEND NHS TRUST	11562030	25,675.00	6100070294
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	ACUTE COMMISINGING	CRAFTFORD AND GRAVESEND NHS TRUST	11562030	25,675.00	6100070294
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	-3,563.11	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11564529	260,000.00	6100081976
Department of Health	NH	Beasey	CGG	10/04/2021	Health Srv Rec Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS			

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

Expenditure Over Threshold

Department of Health	NHS	Beiley	CGG	10/04/2021	Clinical/Medical-Old Public Sector	COMMUNITY SERVICES	BEILEY COUNCIL	11/595633	652,700.75	6100071658
Department of Health	NHS	Beiley	CGG	10/04/2021	Clinical/Medical-Old Public Sector	COMMUNITY SERVICES	BEILEY COUNCIL	11/595662	100,000.00	6100071661
Department of Health	NHS	Beiley	CGG	10/04/2021	Other Gen. Supplies 1-50	COMMUNITY SERVICES	BEILEY COUNCIL	11/595662	11,550.00	6100071674
Department of Health	NHS	Beiley	CGG	10/04/2021	Clinical/Medical-Old Public Sector	WINTER RESIDENCE	BEILEY COUNCIL	11/595662	64,000.00	6100071687
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	CHICAGO COMMUNITY SERVICES	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11/007201	65,250.00	6100070884
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	CHICAGO COMMUNITY SERVICES	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11/007201	65,250.00	6100070884
Department of Health	NHS	Beiley	CGG	10/04/2021	Clinical/Medical-Independent Sector	COMMISSIONING - NON ACUTE	BEILEY COUNCIL	11/16004	30,000.00	6100070989
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	ACUTE COMMISSIONING	NHS HEALTH LTD	11/181704	38,540.00	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	ACUTE COMMISSIONING	NEWCASTLE NHS FOUNDATION TRUST	11/023784	1,000.00	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Clinical/Medical-Voluntary Sector	HOSPICES	GREENWICH & BEILEY COMMUNITY HOSPICE	11/828137	218,000.00	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	ACUTE COMMISSIONING	LEWISHAM & GREENWICH NHS TRUST	11/007201	218,000.00	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PROGRAMME PROJECTS	NHS PROPERTY SERVICES LTD	11/861107	218,740.00	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PROGRAMME PROJECTS	NHS PROPERTY SERVICES LTD	11/861107	218,740.00	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	ACUTE COMMISSIONING	CANTFORD AND GRAVESEND NHS TRUST	11/594232	200,000.00	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	ACUTE COMMISSIONING	CANTFORD AND GRAVESEND NHS TRUST	11/594232	200,000.00	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10/04/2021	Health Svc Res Fdtn Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11/564230	3,560.113.20	6100072662
Department of Health	NHS	Beiley	CGG	10						

Expenditure Over Threshold

Armchair Auditing: What could possibly go wrong?

A3131

Expenditure Over Threshold

Department Name	City	County	Contract Name	Contract Type	Supplier	Invoice Number	Net A/P Amount	Registration number	Business Invoice number
Department of health	Benley	CGO	Clinical/Medical-Clinic Public Sector	COMMUNITY SERVICES	BEXLEY COUNCIL	11595833	466,000.00		6100071665
Department of health	Benley	CGO	Clinical/Medical-Clinic Public Sector	COMMUNITY SERVICES	BEXLEY COUNCIL	11595825	100,000.00		6100071661
Department of health	Benley	CGO	Older Gen Supplies & Inv	COMMUNITY SERVICES	BEXLEY COUNCIL	11595827	15,669.00		6100071669
Department of health	Benley	CGO	Clinical/Medical-Clinic Public Sector	WILLER MANSON	BEXLEY COUNCIL	11595892	64,000.00		6100071687
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11607291	68,226.00	604948867	610008944
Department of health	Benley	CGO	Heath Sln Res 100% Fully Funded	COMMUNITY SERVICES	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11607290	1,000,000.00	61000895	61000895
Department of health	Benley	CGO	Clinical/Medical-Voluntary Sector	COMMUNITY SERVICES - NON ACUTE	BEXLEY COUNCIL	11615604	30,000.00		6100071693
Department of health	Benley	CGO	Clinical/Medical-Independent Sector	ACUTE COMMISSIONING	HEALTHY LIFE FOUNDATION TRUST	11615610	38,528.24	70742405	HL02056882
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	HEALTHY LIFE FOUNDATION TRUST	11615611	58,147.4		HL02056882
Department of health	Benley	CGO	Clinical/Medical-Voluntary Sector	HOSPICES	GREENWICH & BEXLEY COMMUNITY HOSPICE	11615637	218,000.00		6100071677
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	ACUTE COMMISSIONING	LEWISHAM & GREENWICH NHS TRUST	11634001	4,440,144.00		6100071682
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	COMMUNITY SERVICES	NHS PROPERTY SERVICES LTD	11634002	21,517.00		6100071682
Department of health	Benley	CGO	Kent	PROGRAMME PROJECTS	NHS PROPERTY SERVICES LTD	11642864	25,672.31		6100071682
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	ACUTE COMMISSIONING	DARTFORD AND GRAVESEND NHS TRUST	11642865	25,672.31		6100071682
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	ACUTE COMMISSIONING	DARTFORD AND GRAVESEND NHS TRUST	11642866	25,672.31		6100071682
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11654420	-3,662,118.24		CM00218970
Department of health	Benley	CGO	Clinical/Medical-Commercial Sector	COMMUNITY SERVICES	BEXLEY HEALTH LTD	11655629	26,332.45		6100071682
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	AMBUCLANCE SERVICES	LONDON AMBUCLANCE SERVICE NHS TRUST	11663375	587,757.29	604911392	100111154
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11666893	786,836.00		227,489
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11666895	786,836.00		227,489
Department of health	Benley	CGO	Cont Care Adult 100% Fully Funded	FUNDED NURSING CARE	BURA CARE HOMES	11668802	26,812.00	239731641	ALM0369320CALM036938
Department of health	Benley	CGO	Charges From GCU	RECHARGES	NHS SOUTH EAST GCU	11669995	85,249.00	GB054442045	731160994
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11701003	278,828.00		1206094
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	WOOLFELDS EYE HOSPITAL NHS FOUNDATION TRUST	11650115	82,568.00		CM0025903
Department of health	Benley	CGO	Hinew-Foundation Trusts	COMMUNITY SERVICES	OXLEAS NHS FOUNDATION TRUST	11701702	228,988.00		278986
Department of health	Benley	CGO	Hinew-Foundation Trusts	COMMUNITY SERVICES	OXLEAS NHS FOUNDATION TRUST	11701703	141,917.00		278986
Department of health	Benley	CGO	Hinew-Foundation Trusts	COMMUNITY SERVICES	OXLEAS NHS FOUNDATION TRUST	11701710	807,726.00		278986
Department of health	Benley	CGO	Income Mx < 11	BALANCE SHEET	HMRC	11720068	47,713.00	604823331	100200565770 MAR 10
Department of health	Benley	CGO	Income Mx < 11 - Student Loans	BALANCE SHEET	HMRC	11720068	738.00	604823331	100200565770 MAR 10
Department of health	Benley	CGO	National Insurance < 1 - YN-PMS	BALANCE SHEET	HMRC	11720068	29,114.84	604823331	100200565770 MAR 10
Department of health	Benley	CGO	National Insurance < 1 - LAGE	BALANCE SHEET	HMRC	11720068	19,284.18	604823331	100200565770 MAR 10
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11724081	213,245.00		202976
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11724085	268,141.00		202976
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11724086	27,928.00		202976
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11722209	268,424.00		231364
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11724114	224,793.00		202976
Department of health	Benley	CGO	Clinical/Medical-Clinic Public Sector	PLANNED CARE	LEAFLE EYE TRUST	11724087	449.31		202976
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Over Under Performance	ACUTE COMMISSIONING	QUEEN VICTORIA HOSPITAL NHS FOUNDATION TRUST	11738262	-20,596.00		1365
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Over Under Performance	ACUTE COMMISSIONING	WOOLFELDS EYE HOSPITAL NHS FOUNDATION TRUST	11738269	30,540.00		CM002688
Department of health	Benley	CGO	Hinew-Foundation Trusts	COMMUNITY SERVICES	OXLEAS NHS FOUNDATION TRUST	11748400	548,505.00		175451
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	QUEEN VICTORIA HOSPITAL NHS FOUNDATION TRUST	11748386	47,900.00		201009982
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	ACUTE COMMISSIONING	BATES HEALTH UNIT TRUST	11771720	1,816.00		263322
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	ACUTE COMMISSIONING	DARTFORD AND GRAVESEND NHS TRUST	11771723	2,691,942.00		1000041308
Department of health	Benley	CGO	Hinew-Foundation Trusts	MENTAL HEALTH CONTRACTS	SOUTH LONDON AND MAUDSLAY NHS FOUNDATION TRUST	11777827	25,202.00		80007387
Department of health	Benley	CGO	Clinical/Medical-Independent Sector	COMMUNITY SERVICES - NON ACUTE	BEXLEY COUNCIL	11777828	1,000,000.00		6100071682
Department of health	Benley	CGO	Hinew-Foundation Trusts	MENTAL HEALTH CONTRACTS	OXLEAS NHS FOUNDATION TRUST	11782420	1,728,324.00		378977
Department of health	Benley	CGO	Hinew-Foundation Trusts	MENTAL HEALTH CONTRACTS	OXLEAS NHS FOUNDATION TRUST	11801176	30,000.00		378978
Department of health	Benley	CGO	Hinew-Foundation Trusts	ESTATES AND FACILITIES	NHS PROPERTY SERVICES LTD	11801177	51,209.19		484116351
Department of health	Benley	CGO	Hinew-Foundation Trusts	CHILD AND ADOLESCENT MENTAL HEALTH	OXLEAS NHS FOUNDATION TRUST	11801820	258,781.00		378975
Department of health	Benley	CGO	Hinew-Foundation Trusts	CHILD AND ADOLESCENT MENTAL HEALTH	OXLEAS NHS FOUNDATION TRUST	11819430	55,506.00		378974
Department of health	Benley	CGO	Clinical/Medical-Clinic Public Sector	LEARNING DIFFICULTIES	BEXLEY COUNCIL	11819431	21,416.00		6050019263
Department of health	Benley	CGO	Clinical/Medical-Voluntary Sector	HOSPICES	GREENWICH & BEXLEY COMMUNITY HOSPICE	11819432	276,741.00		589
Department of health	Benley	CGO	Clinical/Medical-Commercial Sector	AMBUCLANCE SERVICES	WATFORDS INTERNATIONAL	11836889	25,000.00	882426016	6100071687
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11844000	25,000.00		243430
Department of health	Benley	CGO	Clinical/Medical-Clinic Public Sector	LEARNING DIFFICULTIES	BEXLEY COUNCIL	11844467	250,000.00		6100071247
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11857126	1,171,126.00		6100071682
Department of health	Benley	CGO	Clinical/Medical-Independent Sector	COMMUNITY SERVICES	BELLEFORE ROAD SURGERY	11859391	65,592.39		BA0302016
Department of health	Benley	CGO	Healthcare Sln Res CCG	NHS 111	NHS BROOKLY CCG	11923323	71,747.88		1062702149
Department of health	Benley	CGO	Clinical/Medical-Independent Sector	COMMUNITY SERVICES	KAY-CARD TOWN SURGERY	11923324	36,700.00		6100071682
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11923375	1,137,446.48		234497
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	GUY'S & ST THOMAS HOSPITAL NHS FOUNDATION TRUST	11931340	999,999.99	604923417	2106559
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11933581	-3,103,812.00	604923417	6100071682
Department of health	Benley	CGO	Heath Sln Res NHS Trust-Contract Baseline	PLANNED CARE	LEWISHAM & GREENWICH NHS TRUST	11933581	-129,548.00		8464605
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11951610	1,167,869.00		233736
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11951611	1,167,869.00		CM00218969
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11951612	229,254.29		233340
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11951690	-411,000.00		CM00218970
Department of health	Benley	CGO	Cont Care Adult 100% Fully Funded	CHC AD FULL FUND PERKS HLTH RD	BEXLEY COUNCIL	11951691	269,574.97		6100071682
Department of health	Benley	CGO	Heath Sln Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11959329	68,226.00	604948867	610008944

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.

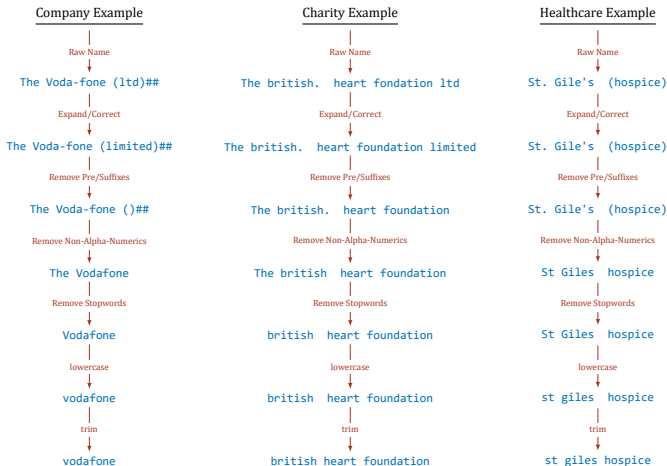
What does a typical ETL pipeline look like in this domain?

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.

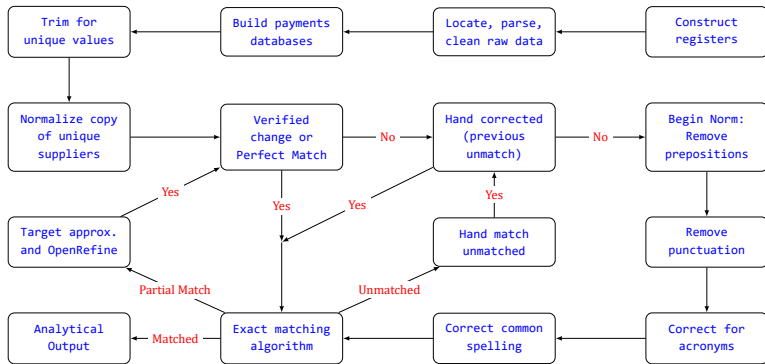
Example Reconciliation Issue

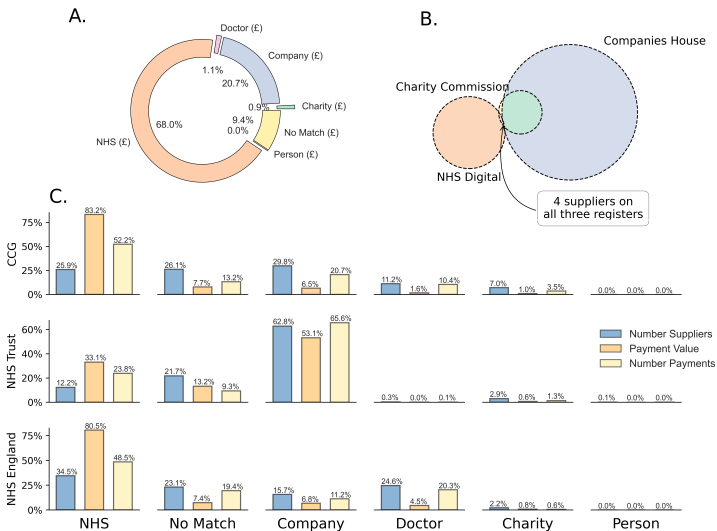
These are two real companies:

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

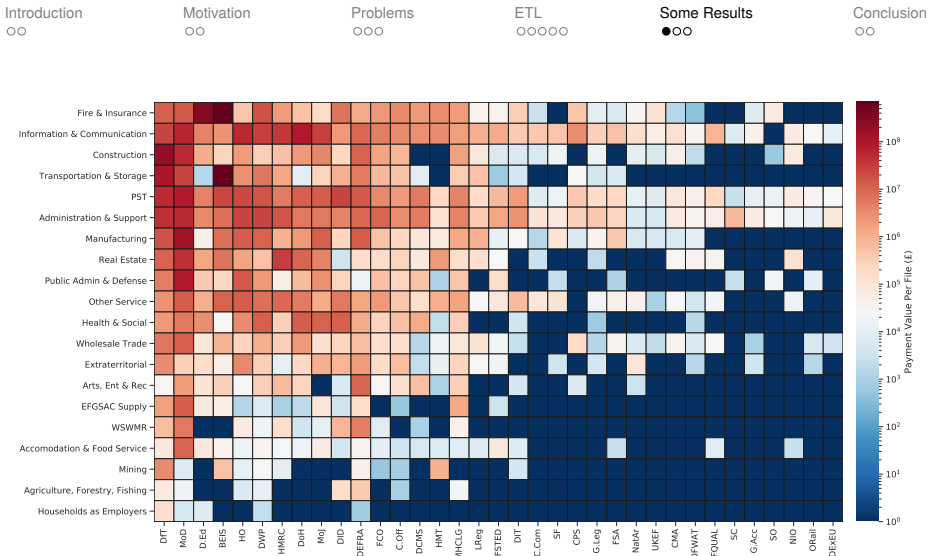
What happens if we normalize names as per the previous slide?

Many Possible Paths to Take

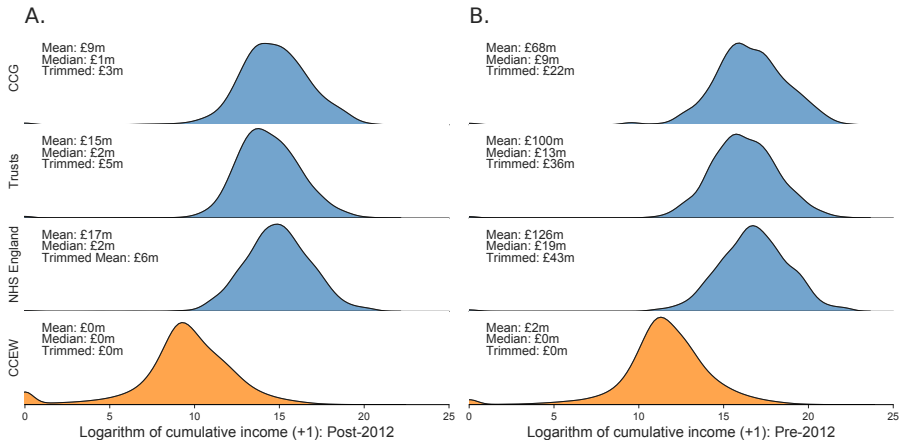




‘What can we match?’



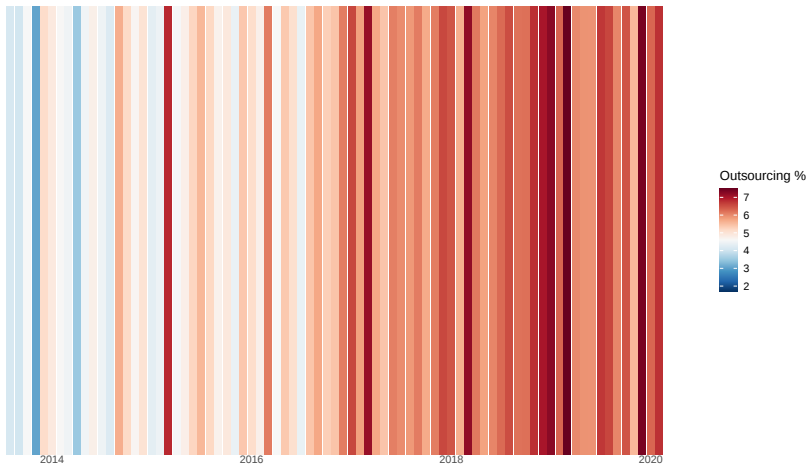
‘Tools for Transparency in Central Government Spending’



‘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!

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/crahal!