



# Data (to fund) Science

Wellcome Data labs

Nick Sorros, 20 November 2018

# Agenda

- Intro to Wellcome
- Policy tool
- Next steps
- Take away points

# Intro to Wellcome

# Me



- MSc Advanced Computing at Imperial college
- Data scientist for 5 years
- Startups like Conversocial, 6tribes and Seedrs
- Communities such as PyData and DataKind
- Senior data scientist in Wellcome Data labs

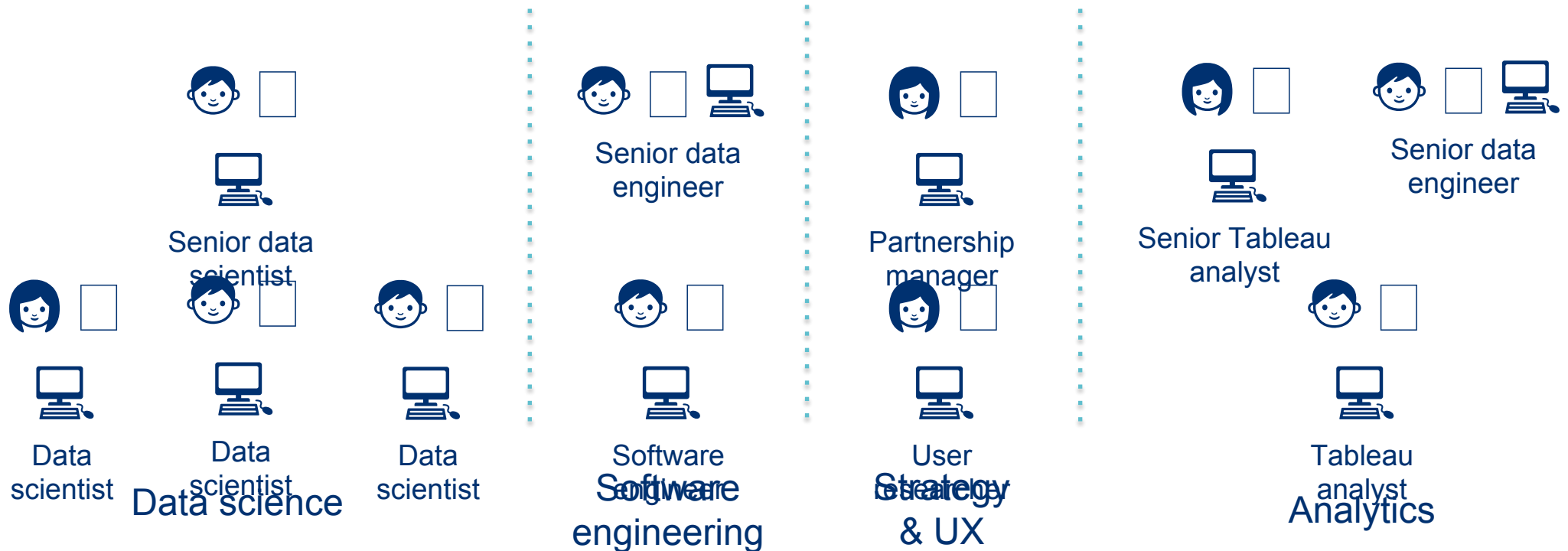
# Wellcome



- 23.2B investment portfolio (2<sup>nd</sup> biggest foundation in the world)
- 2<sup>nd</sup> biggest funder of research in the UK after UKRI
- Funded 1.1B in 2017 across 939 awards
- We fund basic science with a focus on infections diseases and genomics
- Also operate a museum, fund public engagement work and do policy

# Data labs

“Inform Wellcome decision making”

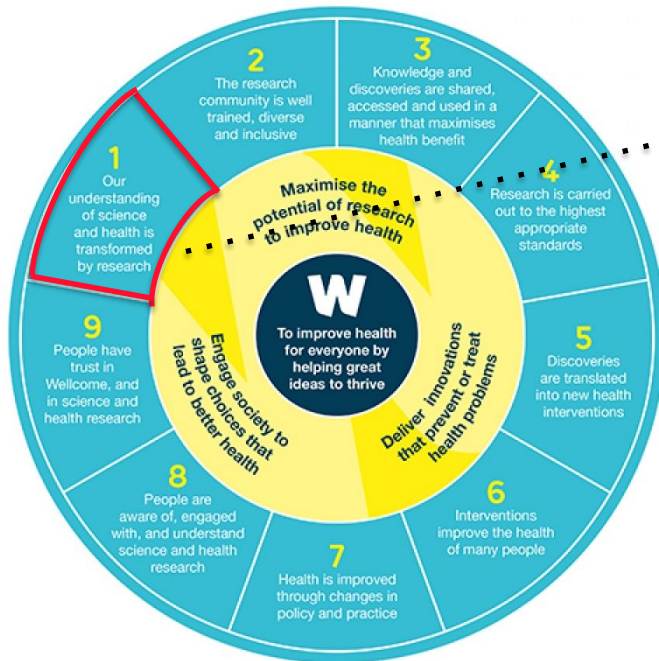


# Mission



To improve health for everyone by helping great ideas to thrive

# Ambition 1



Our understanding of science and health is transformed by research



# Ambition 7



Health is improved through changes in policy and practices

# Policy tool

# Indicators

Health is improved through changes in policy and practices

- a. Policy and practice are informed by research and researchers
  - 1. Number of policy cases that reference WT funded research
  - 2. Number of grantees involved in policy activities
- b. Decision makers take up Wellcome's position on policy and practice
  - 1. Number of Wellcome direct interventions aimed to influence policy

# Manual approach

World Health Organization | iris. Institutional Repository for Information Sharing

English ▾

Search IRIS Search

## Search

All of IRIS ▾  Go

☐ Items with full text online

Now showing items 1-10 of 220886

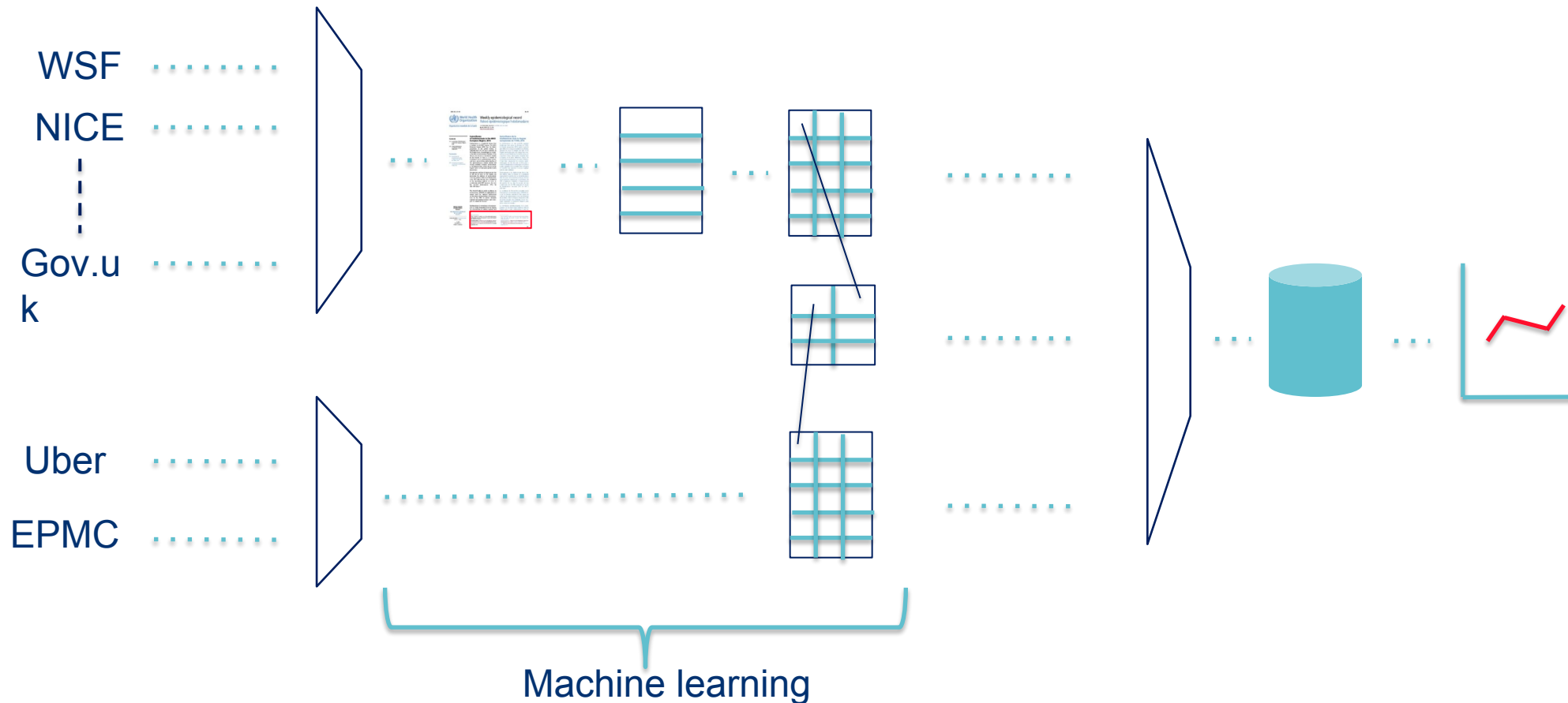
Show Advanced Filters 



**Weekly Epidemiological Record, 2018, vol. 93, 42 [full issue]**  
World Health Organization = Organisation mondiale de la Santé (2018-10-19)

220.886

# Automatic approach



# Scrape

World Health Organization | **iris.**  
Institutional Repository for Information Sharing

English ▾


Search IRIS Search


## Search

All of IRIS ▾  Go

☐ Items with full text online

Show Advanced Filters

Now showing items 1-10 of 220886 



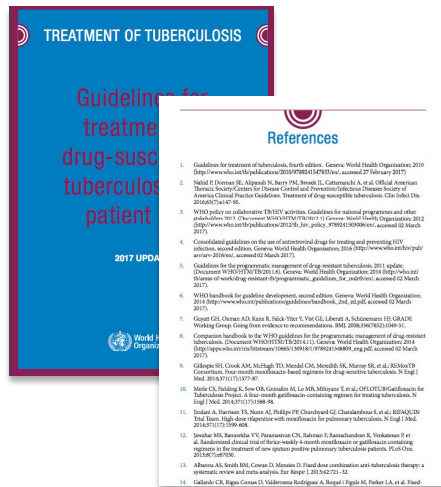
**Weekly Epidemiological Record, 2018, vol. 93, 42 [full issue]**  
World Health Organization = Organisation mondiale de la Santé (2018-10-19)



View/Open

 **WER9346.pdf (1.203Mb)**

# Find



```
{'Reference': ' REFERENCES\n1.\nGDP ranking.[website]. Washington, DC: World\nBank; 2016. (http://data.worldbank.org/\ndata-\ncatalog/GDP-ranking-table, accessed 17\nNovember 2016)\n2.\nLiberia Service Availability and Readiness\nAssess\nment and Quality of Care. [Draft\nDocument]. Government of Liberia; October\n2016\n3.\nUnited Nations, World Bank, Eu\nropean Union ,\nAfrican Development Bank. Recovering from\nthe Ebola Crisis. New York: United Nations\nDevelopment Pr\nogramme; 2015\n(http://www.undp.org/content/undp/en/home/lib\nnrarypage/crisis-prevention-and-\nrecovery/recovering-fr\nom-the-ebola-crisis---full-\nreport.html, accessed 8 November 2015)\n4.\nMinistry of Health and Social Welfare Annual\nReport 2014. Monrovia: Republic of\nLiberia;2014\n(http://reliefweb.int/sites/reliefweb.int/files/reso\nnurses/MOHS\nW%20Annual%20Report%202014\nRevised.pdf, accessed 8 November 2016)\n5.\nMinistry of Health and Social Welfare. Natio\nnal\nHealth and Social Welfare Policy and Plan\n(2011-2021). Monrovia: Republic of Liberia ;\n2011\n(http://www.mohs\nw.gov.lr/documents/Final%20NHPP%20(high%20res\n).pdf, accessed 8 November 2016)\nhttps://www.healthresearchweb.org/f\niles/NHPP\nJuly132011.pdf\n6.\nInvestment Plan for Building a Resilient Health\nSystem in Liberia 2015 to 2021. Monr\novia:\nRepublic of Liberia; 2015\n(http://pages.au.int/sites/default/files/LIBERIA-\nInvestment%20Plan%20for%20Bui\nlding%20a%20Resilient%20Health%20System.pdf,\naccessed 8 November 2016)\n7.\n2008 National Population and Housing C\nensus:\nPreliminary Results. Monrovia: Republic of\nLiberia; 2008\n45\nReferences\n(unstats.un.org/unsd/dnss/docViewe\nr.aspx?docID=2075, accessed 8 November 2016)\n8.\nAfrican Partnerships for Patient Safety. Patient\nSafety Situatio\nnal Analysis ( Short Form).\nGeneva: World Health Organization; 2012\n(http://www.who.int/patientsafety/implementation/\nnon/apps/resources/APPS_Improv_PS_Situational\nAnalysis_SF_2012_04_EN.pdf, accessed 8\nNovember 2016)\n9.\nHand Hyg\niene Self-Assessment Framework\n2010. Geneva: World Health Organization;\n2010\n(http://www.who.int/gpsc/country_wor\nk/hhsa_fr\namework October 2010.pdf?ua=1. accessed 8\nNovember 2010)\n10.\nPatient Safetv. Copenhagen: WHO Regional\n
```

# Split

```
[ ' REFERENCES',  
  'GDP ranking.[website]. Washington, DC: World Bank; 2016. (http://data.worldbank.org/datacatalog/GDP-ranking-table,  
  accessed 17 November 2016)',  
  'Liberia Service Availability and Readiness Assessment and Quality of Care. [Draft Document]. Government of Liberia;  
  October 2016',  
  'United Nations, World Bank, European Union , African Development Bank. Recovering from the Ebola Crisis. New York:  
  United Nations Development Programme; 2015 (http://www.undp.org/content/undp/en/home/librarypage/crisis-prevention-andrecovery/recovering-from-the-ebola-crisis---fullreport.html, accessed 8 November 2015)',  
  'Ministry of Health and Social Welfare Annual Report 2014. Monrovia: Republic of Liberia;2014 (http://reliefweb.int/sites/reliefweb.int/files/resources/MOHSW%20Annual%20Report%202014\_Revised.pdf, accessed 8 November 2016)',  
  'Ministry of Health and Social Welfare. National Health and Social Welfare Policy and Plan (2011-2021). Monrovia: Republic of Liberia ; 2011 (http://www.mohsw.gov.lr/documents/Final%20NHPP%20\(high%20res\).pdf; accessed 8 November 2016) https://www.healthresearchweb.org/files/NHPP\_July132011.pdf',  
  'Investment Plan for Building a Resilient Health System in Liberia 2015 to 2021. Monrovia: Republic of Liberia; 2015 (http://pages.aun.int/sites/default/files/LIBERIA%20Investment%20Plan%20for%20Building%20a%20Resilient%20Health%20System.pdf; accessed 8 November 2016)',  
  '2008 National Population and Housing Census: Preliminary Results. Monrovia: Republic of Liberia; 2008 45 References (unstats.un.org/unsd/dnss/docViewer.aspx?docID=2075, accessed 8 November 2016)',  
  'African Partnerships for Patient Safety. Patient Safety Situational Analysis ( Short Form). Geneva: World Health Organization; 2012. (http://www.who.int/patientsafety/implementation/apps/resources/APPS\_Improv\_PS\_Situational\_Analysis\_SF\_2012\_04\_EN.pdf, accessed 8 November 2016)',  
  'Hand Hygiene Self-Assessment Framework 2010. Geneva: World Health Organization; 2010 (http://www.who.int/gpsc/county\_work/hhsa\_framework\_October\_2010.pdf?ua=1, accessed 8 November 2010)',  
  'Patient Safety. Copenhagen: WHO Regional Office for Europe World Health Organization; 2016 (http://www.euro.who.int/en/healthtopics/Health-systems/patient-safety, accessed 21 November 2016)',  
  'A Guide to the Implementation of the WHO Multimodal Hand Hygiene Improvement Strategy. Geneva: World Health Organization; 2009 (http://www.who.int/gpsc/5may/Guide\_to\_Impl (http://www.who.int/gpsc/5may/Guide\_to\_Impl World Health Organization 20 Avenue Appia CH-1211 Geneva 27 Switzerland Tel.: +41 22 791 5060 Email: abramahmn@who.int Please visit us at: Please visit us at:']
```

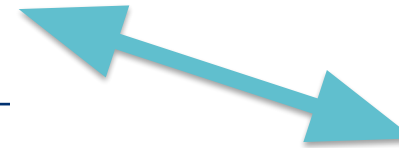


# Parse

| Document | Reference number | Authors | Issue       | Journal                        | Pagination | PubYear | Title   | Volume | PubYear Cleaned |
|----------|------------------|---------|-------------|--------------------------------|------------|---------|---|--------|-----------------|
| 0        | 5                | 2       |             |                                |            |         | Liberia Service Availability and Readiness Ass... |        |                 |
| 1        | 5                | 3       |             | United Nations, European Union |            |         | African Development Bank, Recovering from the ... |        |                 |
| 2        | 5                | 4       |             |                                |            |         | Ministry of Health and Social Welfare Annual R... |        |                 |
| 3        | 5                | 5       |             |                                |            |         | lr/documents/Final%20NHPP%20(high%20res)          |        |                 |
| 4        | 5                | 6       |             |                                |            |         | Investment Plan for Building a Resilient Healt... |        |                 |
| 5        | 5                | 7       | docl D=2075 |                                |            |         | 2008 National Population and Housing Census: P... |        |                 |
| 6        | 5                | 8       |             |                                |            |         | African Partnerships for Patient Safety, Patie... |        |                 |
| 7        | 5                | 9       |             |                                |            |         | Hand Hygiene Self-Assessment Framework 2010       |        |                 |
| 8        | 5                | 10      |             |                                |            |         | Copenhagen: WHO Regional Office for Europe Wor... |        |                 |

# Match

| Predicted authors          | Predicted title   | Predicted journal |
|----------------------------|---|-------------------|
| Hong Chau TT, Hoang Mai HT | Timing of initiation of antiretroviral therapy in human immunodeficiency virus (HIV)-associated tuberculosis meningitis |                   |
| Rahman A, Rwagatare P      | Grand challenges: integrating maternal mental health into maternal and child health programmes                          | PLoS Med          |



## Dimensions title

Timing of Initiation of Antiretroviral Therapy in Human Immunodeficiency Virus (HIV) - Associated Tuberculous Meningitis

38. Torok ME, Bich Yen NT, Hong Chau TT, Hoang Mai HT, Phu NP, et al. Timing of initiation of antiretroviral therapy in human immunodeficiency virus (HIV)-associated tuberculosis meningitis. Clin Infect Dis. 2011;52(11):1374–83.

# Accuracy

Find



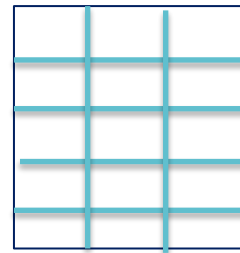
85%

Split



57%

Parse



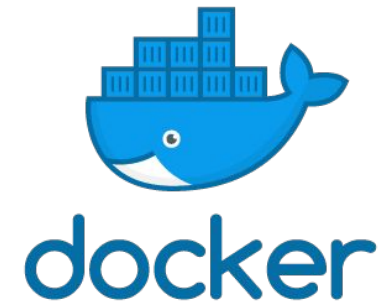
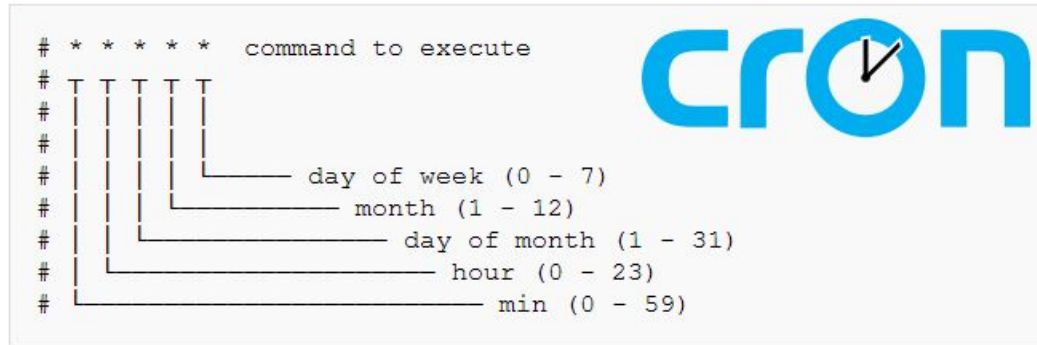
87%

Match

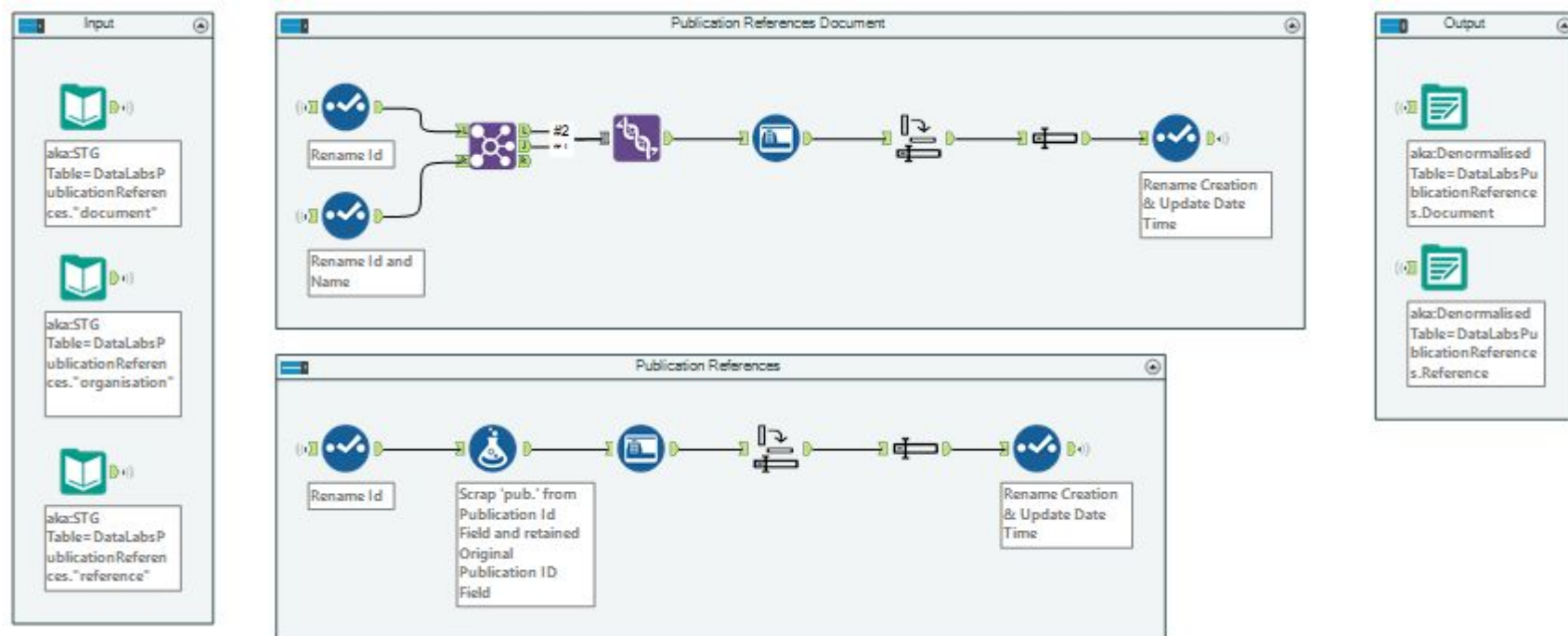


96%

# Orchestration

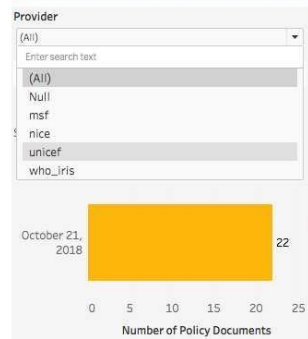


# ETL



# Dashboard

## Policy document - WT reference matches

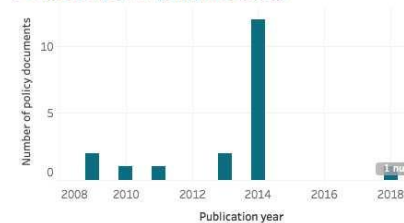


56 unique policy document - WT reference matches

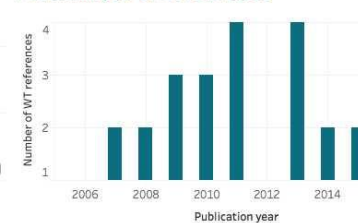
31 policy documents with at least one WT reference

56 unique WT references in the policy documents

## Publication year of policy documents



## Publication year of WT references

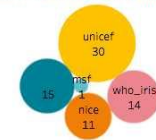


## Searching for keywords in the full text of the policy documents

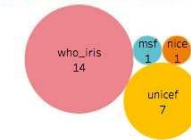
Search Word in Policy Text

the

Total number of policy document - WT reference matches



Number of policy document - WT reference matches with search terms



Search results (click for link to policy document)

|          |   |
|----------|---|
| msf      | Null  |
| nice     | Policy document: Preventing suicide in community and custodial settings2018.<br>WT reference: Hawton K, Linsell L (2014). Self-harm in prisons in England and Wales: an epidemiological study of prevalence, risk factors, clustering, and subsequent suicide. The Lancet   |
| unicef   | Null  |
| who_iris | Policy document: Ending preventable child deaths from pneumonia and diarrhoea by 2025 : the integrated global action plan for pneumonia and diarrhoea (GAPD)2013.<br>WT reference: Chopra M et al (2007). Case Management of Childhood Pneumonia in Developing Countries, The Pediatric Infectious Disease Journal  |
|          | Policy document: Strengthening vital statistics systems : what are the practical interventions necessary to reduce ignorance and uncertainty about causes of death and disease burden in the Asia Pacific region2014. ...<br>Policy document: Strengthening vital statistics systems : what are the practical interventions necessary to reduce ignorance and uncertainty about causes of death and disease burden in the Asia Pacific region2014. ...<br>Policy document: WHO South-East Asia Journal of Public Health, Volume 3, Issue 2, April-June 20142014.<br>WT reference: Hoque DE, Rahman M (2009). Cost Effectiveness in Low- and Middle-income Countries. Pharmacoeconomics<br>Policy document: WHO South-East Asia Journal of Public Health, Volume 3, Issue 2, April-June 20142014.<br>WT reference: Hoque DE, Rahman M (2013). Brucellosis in low-income and middle-income countries. Current Opinion in Infectious Diseases<br>Policy document: WHO South-East Asia Journal of Public Health, Volume 3, Issue 2, April-June 20142014.<br>WT reference: Hoque DE, Rahman M (2015). HIV and the Millennium Development Goals. Archives of Disease in Childhood |

# Volume

| Organisation | Documents | WT References |
|--------------|-----------|---------------|
| Gov.uk       | 102.903   |               |
| Parliament   | 18.744    |               |
| WHO          | 7.484     | 867           |
| NICE         | 289       | 5             |
| Unicef       | 189       | 33            |
| MSF          | 69        | 1             |

# Internal interest

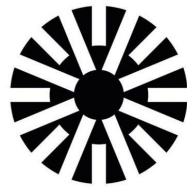
- Science team (RC2H review)
- Policy team (Board of governors report)
- Central team (Horizon scanning)



# External interest

The logo for the Bill & Melinda Gates Foundation, featuring a solid red square background with the text "BILL & MELINDA GATES foundation" in white, serif, all-caps font.

BILL & MELINDA  
GATES foundation



Pew Research Center



World Health  
Organization

UK Research  
and Innovation

# Open source

**wellcometrust / policytool** Watch 6 Unstar 3 Fork 0

Code Issues 16 Pull requests 3 Actions Projects 0 Wiki Insights Settings

Wellcome tool to parse references scraped from policy documents using machine learning [Edit](#)

[Manage topics](#)

277 commits 8 branches 0 releases 5 contributors

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

| nsorros Rename variables, functions and file in separate |   | Latest commit 9f3e764 5 days ago |
|--|---|----------------------------------|
| model_tests  | Also changes the variable names                                 | 3 months ago                     |
| reference_parser_models                                  | new name for models folder                                      | 3 months ago                     |
| tests  | fixing text predict to work with new structure                  | 7 days ago                       |
| utils  | Rename variables, functions and file in separate                | 4 days ago                       |
| wsf-scraper-web  | Merge pull request #54 from wellcometrust/remove-scraper-pipenv | 4 days ago                       |
| .dockerignore  | Dependencies and docker related changes                         | 4 months ago                     |
| .gitignore   | Add a gitignore file  | 4 months ago                     |
| CONTRIBUTING.md  | Add Contributing guidelines and PR template to the repo (#96)   | 4 months ago                     |
| Dockerfile   | Lighter base image  | 3 months ago                     |
| Makefile   | rename virtualenv to follow guidelines                          | 26 days ago                      |

# Next steps

# Improve Split

Find



85%

Split

A diagram of a table with 4 rows and 1 column, representing the result of splitting the table from the 'Find' step. The table is outlined in blue and has a red border around it.

57%

Parse

A diagram of a table with 4 rows and 3 columns, representing the result of parsing the table from the 'Split' step. The table is outlined in blue.

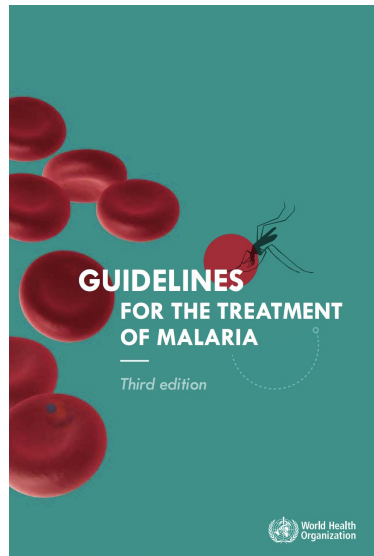
87%

Match

A diagram of a table with 2 rows and 1 column, representing the result of matching the table from the 'Parse' step. The table is outlined in blue.

96%

# Tags



Malaria



## HIV DRUG RESISTANCE REPORT 2017



HIV

# Orchestration

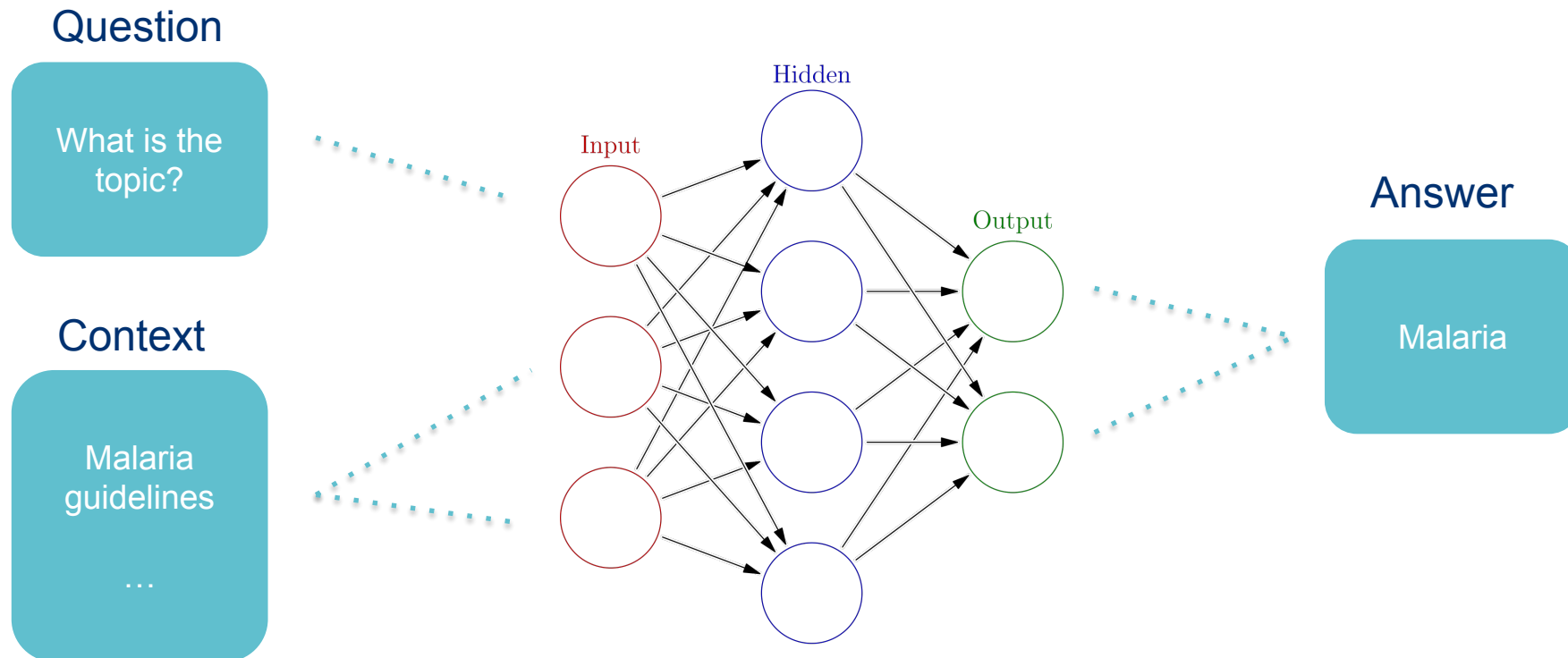


Airflow



Kubernetes

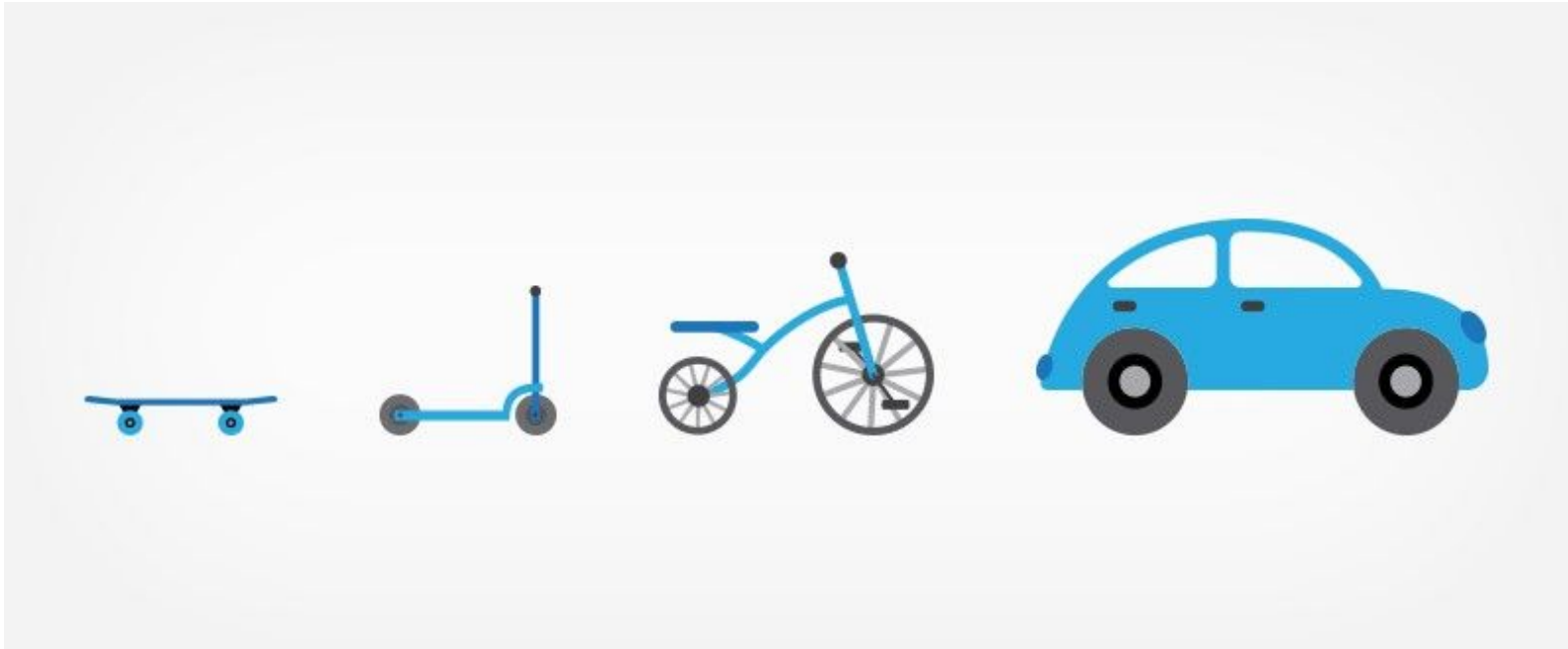
# Neural parser



# Take away points



# Simple before complex



# Challenge using ML



# Create **supply**, not demand



# Questions?

Email: [n.sorros@wellcome.ac.uk](mailto:n.sorros@wellcome.ac.uk)

We are hiring!