The NPD: an open methodology

Peter Kemp
University of Roehampton

@peterejkemp

University of Roehampton London

A bit of background

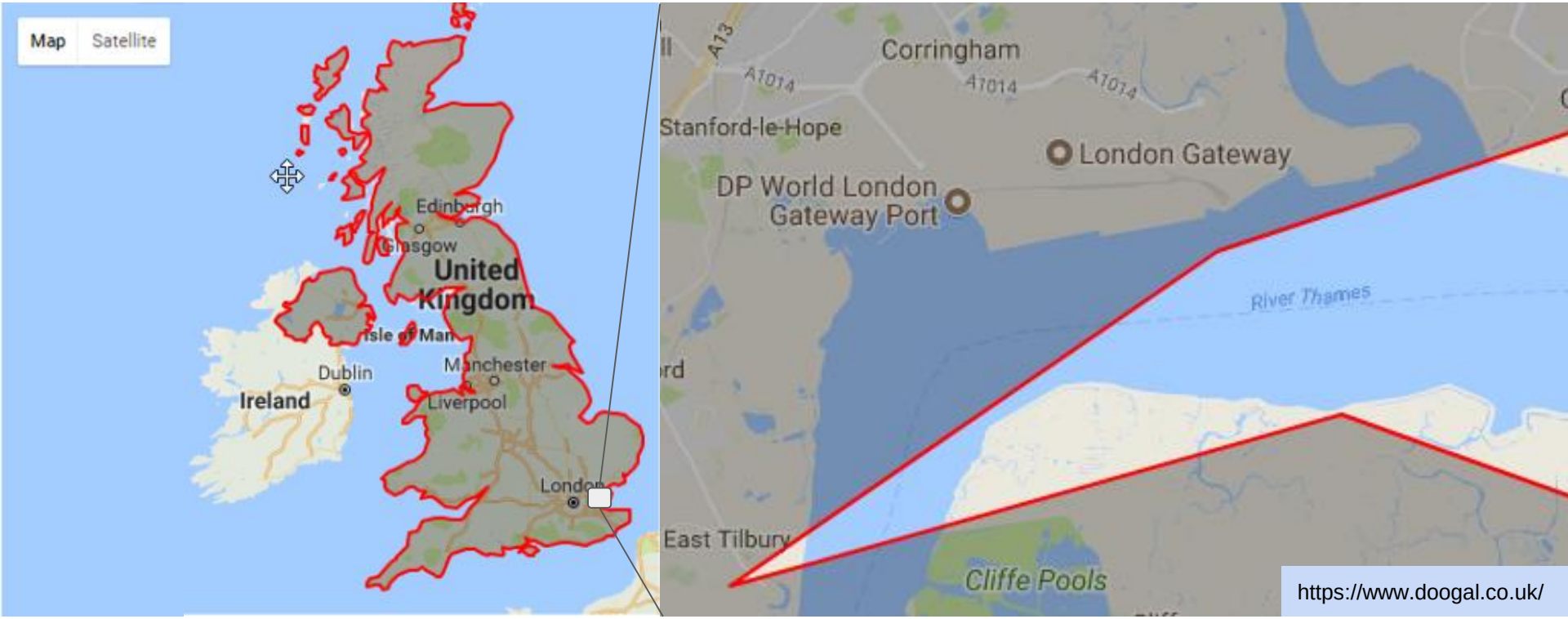
- Teacher trainer with CS background
- 2 years using the NPD
- KS4 & KS5 analysis on computing education in England
- Using R

University of Roehampton London

Problems with research (in general)

University of Roehampton London

A lack of shared definitions



https://www.doogal.co.uk/

5000m? 5500m? Curvature of the earth? High tide?
Working class students

http://www.southendtimeline.com/crowstone.htm

University of Roehampton London

Can I ***easily*** repeat / build on this research?

The impact of Teach First on pupil attainment at age 16

Rebecca Allen, Jay Allnutt

First published: 19 May 2017 Full publication history

DOI: 10.1002/berj.3288 View/save citation

Cited by (CrossRef): 0 articles Check for updates Citation tools

http://onlinelibrary.wiley.com/doi/10.1002/berj.3288/full

Which are the most difficult subjects at GCSE?

By Dave Thomson

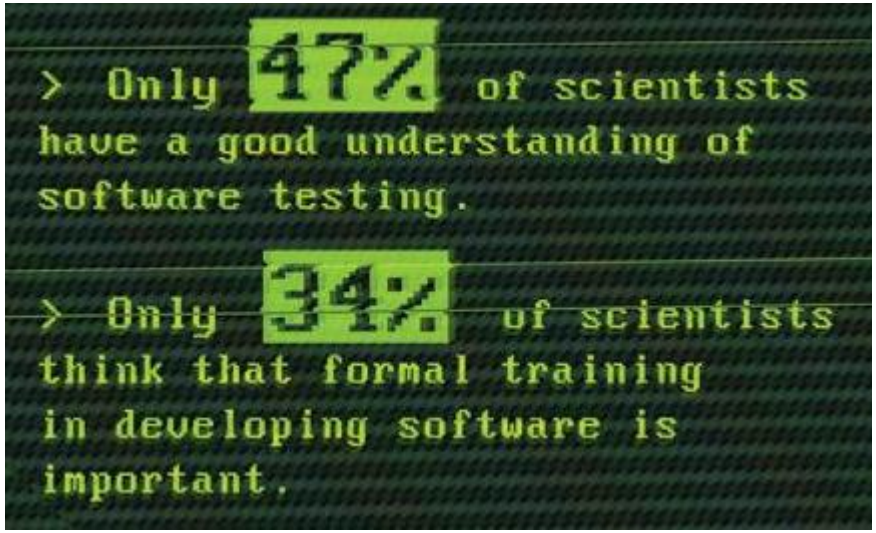
D&T Graphic Products	-0.73	31708
D&T Systems & Control	-0.85	2969
Spanish	-0.99	84846
Classical Civilisation	-1.10	3971
Psychology	-1.21	15897
Computer Studies/Computing	-1.28	33378
Economics	-1.29	9414
French	-1.34	150702
Humanities: Single		

https://educationdatalab.org.uk/2016/02/which-are-the-most-difficult-subjects-at-gcse/

University of Roehampton London

Coding errors and research

In 2006, the team realized that a computer program supplied by another lab had flipped a minus sign, which in turn reversed two columns of input data, causing protein crystal structures that the group had derived to be inverted. Chang says that the other lab provided the code with the best intentions, and **"you just trust the code to do the right job"**. His group was forced to retract five papers published in *Science*, the *Journal of Molecular Biology* and *Proceedings of the National Academy of Sciences*, and now triple checks everything, he says.



http://www.nature.com/news/2010/101013/full/467775a.html

A few problems with the NPD

University of Roehampton London

Non-normalised database tables

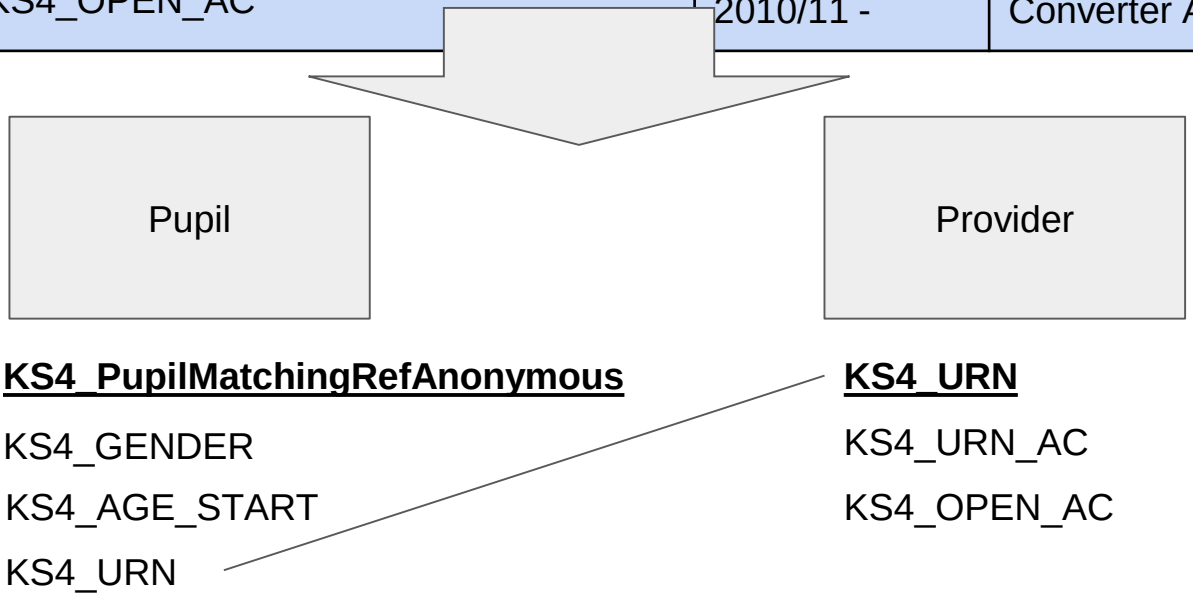
KS4_PupilMatchingRefAnonymous	2001/02 -	Pupil matching reference - Anonymous.
KS4_GENDER	2001/02 -	Gender.
KS4_AGE_START	2001/02 -	Age of pupil at start of the academic year (in full years).
KS4_URN	2001/02 -	School's Unique Reference Number
KS4_URN_AC	2010/11 -	Converter Academy: URN
KS4_OPEN_AC	2010/11 -	Converter Academy: open date

1151 different fields
~700 in my 2015 snapshot

University of Roehampton London

Non-normalised database tables

KS4_PupilMatchingRefAnonymous	2001/02 -	Pupil matching reference - Anonymous.
KS4_GENDER	2001/02 -	Gender.
KS4_AGE_START	2001/02 -	Age of pupil at start of the academic year (in full years).
KS4_URN	2001/02 -	School's Unique Reference Number
KS4_URN_AC	2010/11 -	Converter Academy: URN
KS4_OPEN_AC	2010/11 -	Converter Academy: open date



University of Roehampton London

Missing values

Even though the KS4 students are the same students at KS5 a lot of their demographic data falls away:

```
>> healthOfField(Students_GCSE_12, “EthMaj”)

>> 0.1012753

>> healthOfField(Students_Alevel_14_RAW, “EthMaj”)

>> 0.5169083

>> healthOfField(Students_Alevel_14_matched, “EthMaj”)

>> 0.1900806
```

University of Roehampton London

Inconsistent naming

```
>> EthnicGroupMajor_SPR12

>> EthnicGroupMajor_SPR13

>> EthnicGroupMajor_SPR14

>> EthnicGroupMajor_SPRnn
```

Leading to difficulties in:

- Importing data into your model
- Year on year analysis

I’m sure you have your own pet hates....

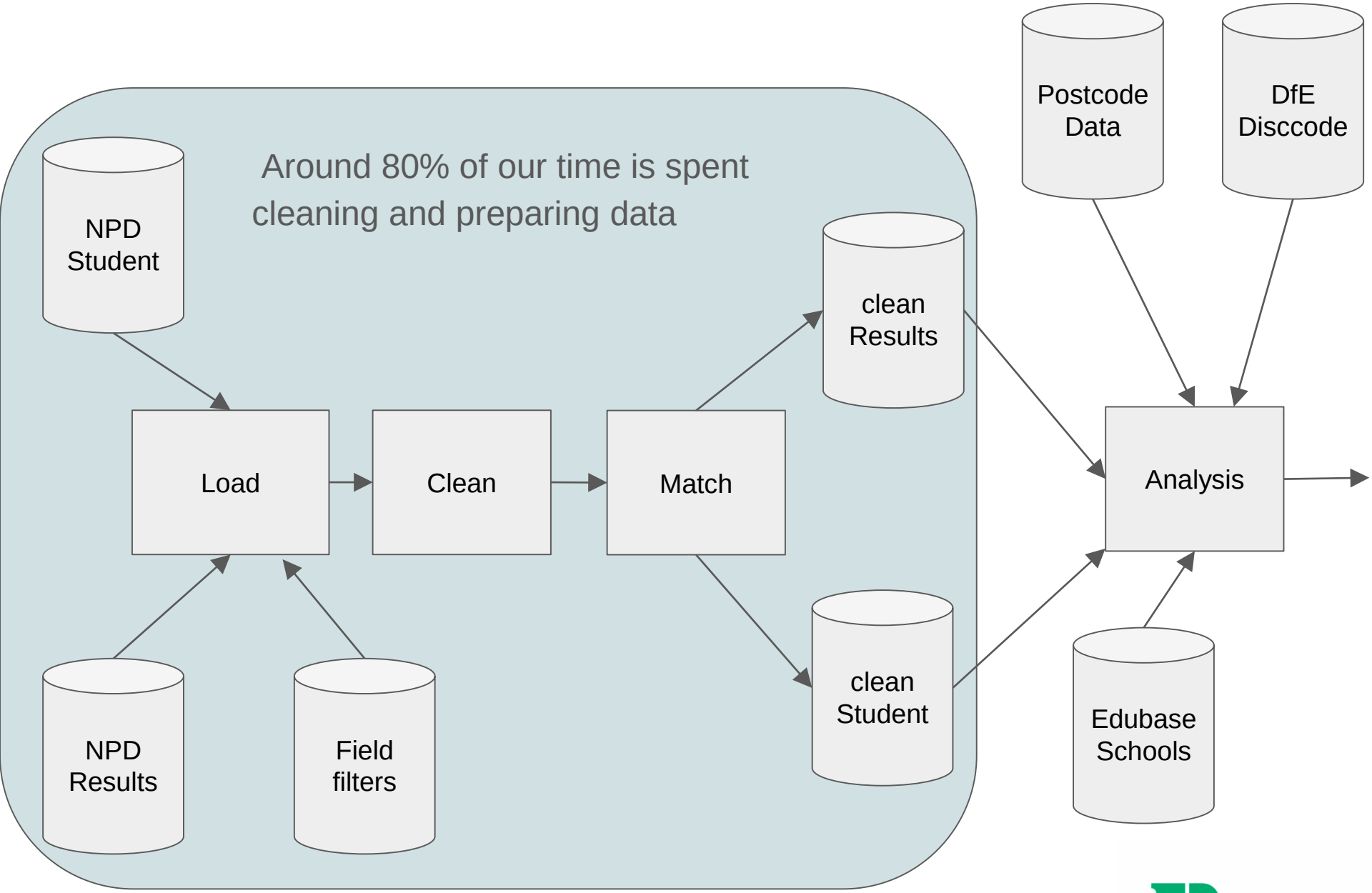
University of Roehampton London

A few problems with NPD analysis

University of Roehampton London

```r
#return a dataframe with the IDs, Names and n of the top 10 largest subjects
#include `subject` if it isn't present
buildSubjectList <-function(spreadResults, subject="X2610", subsize){

  #spreadResults <- Aresults15
  #get columns starting with X followed by at least one number (it denotes a subject ID)
  IDs <- sapply(spreadResults[ , grep("X[0-9]", colnames(spreadResults))],
                function(x) length(which(!is.na(x))) )


  IDs <- cbind(read.table(text = names(IDs)), IDs, row.names = NULL)
  colnames(IDs) <- c("ID", "n")
  #order by largest subjects
  IDs <- IDs %>% arrange(desc(n))

  #fetch subject names
  filename <- paste0(folder,"MappingCodes.txt")
  Mappings <- read.csv(filename, head=TRUE, sep="\t")
  Mappings$MAPPING <- paste0("X",Mappings$MAPPING)

  #TODO: really really really want to put this in a single lapply
  IDs$SubjectName <- unlist(
    lapply(IDs$ID,
           function(x)
             droplevels(Mappings[which(Mappings$MAPPING == x), ]$MAPPING_DESCRIPTION)))

  if(subject %in% IDs[1:subsize,]$ID){
    return(IDs[1:subsize,])
  }else{
    return(rbind(IDs[1:subsize-1,], IDs[c(IDs$ID == subject),]))
  }
}
```

University of Roehampton London



University of Roehampton London

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning* (Vol. 479). John Wiley & Sons.

A solution

University of Roehampton London

TRACER codebase on Github

"given enough eyeballs, *all bugs are shallow*"

Raymond, E. (1999). The cathedral and the bazaar. *Philosophy & Technology*, *12*(3), 23.

“Happy families are all alike; every unhappy family is unhappy in its own way.”

Tolstoy, L. Anna Karenina

"Tidy datasets are all alike, but every messy dataset is messy in its own way."

Wickham, H., & Grolemund, G. (2016). R for data science.

- Free
- Open source
- Multi purpose shared code base for analysis

University of Roehampton London

This repository | Search | Pull requests | Issues | Marketplace | Explore

plukethep / TRACER

Unwatch 1 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

R project to standardise the analysis of the English National Pupil Database

Edit

Add topics

32 commits | 1 branch | 0 releases | 1 contributor

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

plukethep Update Main.R — Latest commit c9caac6 a day ago

code	Update Main.R	a day ago
data	Update help.txt	a day ago
outputs	Create help.txt	a day ago
reports	Update help.txt	a day ago
DfECompareSchools.r	Update DfECompareSchools.r	a month ago
README.md	Initial commit	8 months ago

University of Roehampton London

Separation of data preparation and analysis

3.4.1 GCSE

GCSE computing provision in 2015 shows that the percentage of coastal institutions offering the subject is in line with inland providers. This suggests that both types of institution are able to staff computing courses at roughly the same level.

Table 15: 2015 GCSE computing provision in coastal schools

Type	Total Schools	Total Students	Subject Providers	Providers %	Subject Students	Students %	Average Cohort Size
Inland	4191	497261	1188	28.3	27439	5.5	23.1
Coastal	844	99466	245	29.0	5385	5.4	22.0

Your report.rmd

```
882 ###GCSE
883 GCSE `r subject` provision in `r year` shows that the percentage of coastal
    institutions offering the subject is in line with inland providers. This
    suggests that both types of institution are able to staff computing courses at
    roughly the same level.
884
885 ```{r GCoastalTable, results='asis'}
886 temp <- OutputCoastalSchoolSubjectbreakdown(GresultsCURRENT, "X2610", "GCSE")
887
888 #get rid of the NAs
889 temp <- temp %>% filter(!is.na(Type)) %>% mutate(Type = ifelse(!Type, "Inland",
    "Coastal"))
890
```

TRACER codebase

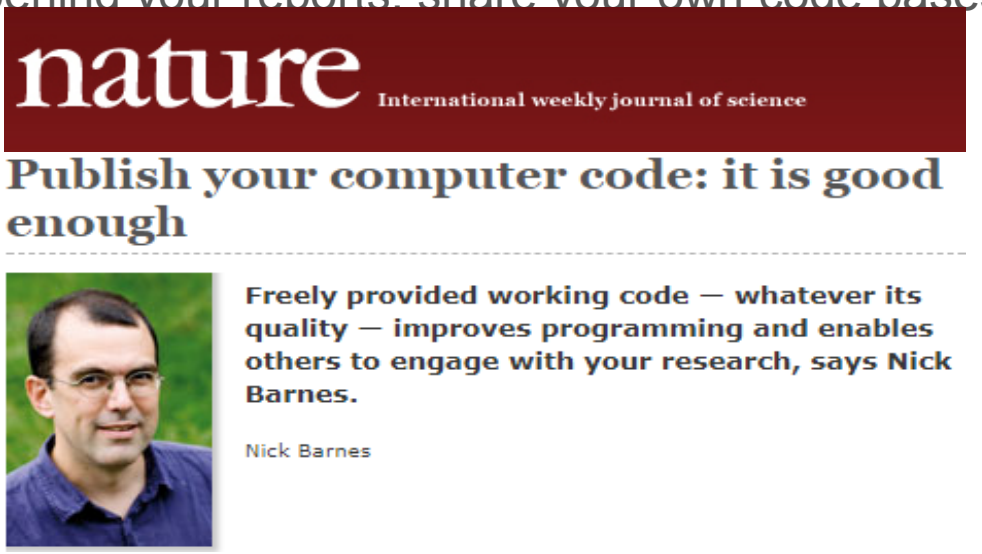
University of Roehampton London

What to do now

Look at: https://github.com/plukethep/TRACER

- Be critical but constructive
- Suggest and make changes

Publish the code behind your reports, share your own code bases, it doesn't have to be R...



nature International weekly journal of science

Publish your computer code: it is good enough

Freely provided working code — whatever its quality — improves programming and enables others to engage with your research, says Nick Barnes.

Nick Barnes

https://www.nature.com/news/2010/101013/full/467753a.html

University of Roehampton London