



R in the Wild: A Safari Adventure through Industry

Nic Crane

Data Scientist

✉ ncrane@mango-solutions.com



Agenda

Mango Solutions

The Growth of R in Industry

Methods of Deploying R

Case Studies

Overcoming
Misconceptions



Who are Mango?

- Providers of Data Science Software and Services
- Premier suppliers of R services
 - Also, Python and Spark
- Private company founded in 2002
- Headquartered in the UK, operating globally
- Team of ~65



User Groups

- Proud sponsors of Oxford R!
- We run a number of R user groups:
 - BaselR
 - ManchesterR
 - LondonR
- But have been involved with many others too!



EARL

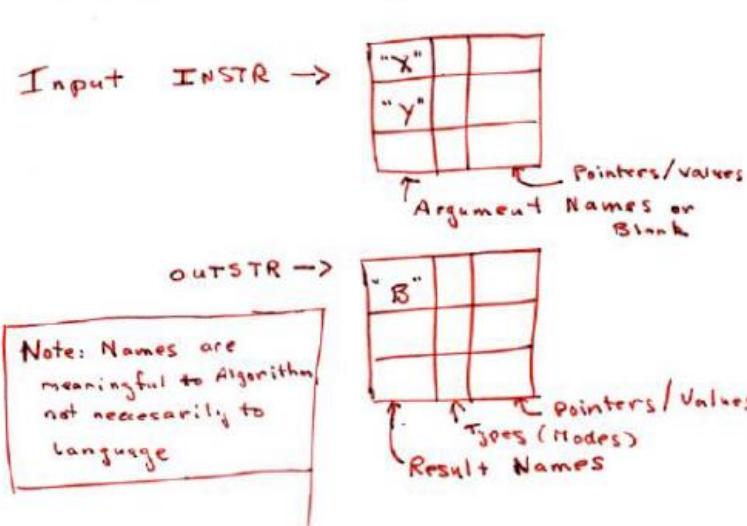
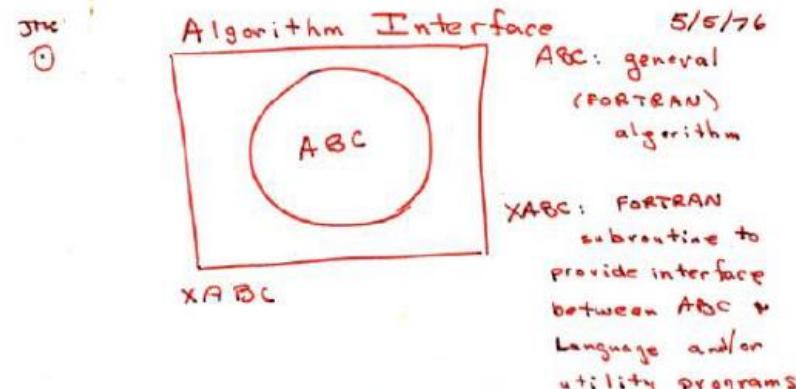
- Enterprise Applications of the R Language
- Every September in London
- You can submit an abstract to speak about anything related to the commercial use of R.
- If you don't want to speak keep an eye out for early bird tickets.



The Rise of R



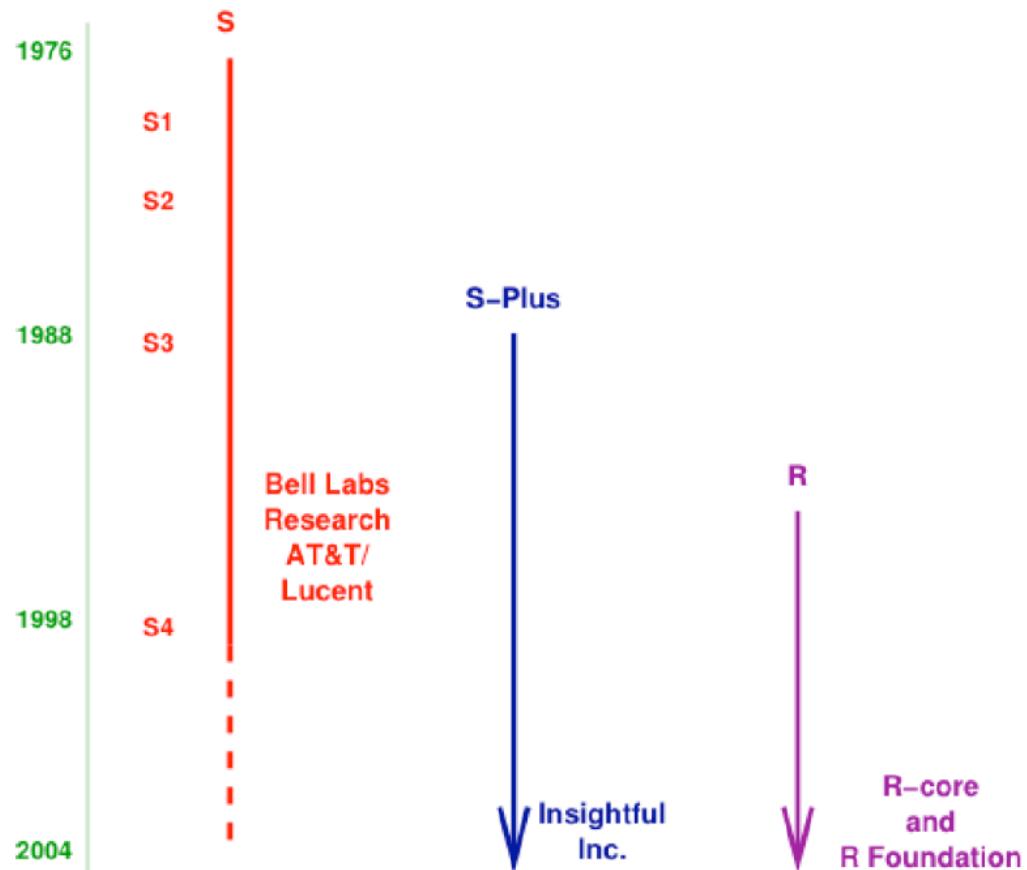
Origins



The best thing about R is that it was written by statisticians. The worst thing about R is that it was written by statisticians.

Bow Cowgill, Google





Old SkoolS \R

- Dealing with Factors
- Creating strings then using eval(parse(text = ...))
- Dollars and Square Brackets
- The apply family of functions
- Trying to get your package to build
- Adding C/C++ code to packages before RCpp
- The assignment character
- Unfriendly Rhelp conversations



Modern R



R Studio Community

[all categories ▶](#) [all tags ▶](#) [Categories](#) [Latest](#) [New \(19\)](#) [Uni](#)

Category	Topics
 tidyverse This category is for anything and everything about the tidyverse.	9 / week 1 new
 RStudio IDE This category is for discussing the RStudio IDE, both desktop and server versions.	9 / week 2 new
 Teaching For discussions about teaching.	2 / week 1 new
 shiny Please ask your questions about shiny here.	19 / week 5 new



[Rd] Milestone: 1000 packages on CRAN as of today(?)

Henrik Bengtsson hb at stat.berkeley.edu

Thu Apr 12 18:09:21 CEST 2007

- Previous message: [\[Rd\] printf capture](#)
- Next message: [\[Rd\] Milestone: 1000 packages on CRAN as of today\(?\)](#)
- **Messages sorted by:** [\[date \]](#) [\[thread \]](#) [\[subject \]](#) [\[author \]](#)

Hi,

I was just looking at the "CRAN Daily Package Check Results"
[\[http://cran.r-project.org/src/contrib/checkSummary.html\]](http://cran.r-project.org/src/contrib/checkSummary.html), and
realized there are 1000 packages on CRAN as of today (look at row 3 in
the table below). Yet another quite extraordinary milestone in R
history.

Last updated on 2007-04-12 11:48:32

Results for installing and checking packages using the three current
flavors of R on systems running Debian GNU/Linux testing (r-devel
ix86: AMD Athlon(tm) XP 2400+ (2GHz), r-devel x86_64: Dual Core AMD
Opteron(tm) Processor 280, r-prerel/r-release: Intel(R) Pentium(R) 4
CPU 2.66GHz), MacOS X 10.4.7 (iMac, Intel Core Duo 1.83GHz), and
Windows Server 2003 SP2 (32-bit) (AMD Athlon64 X2 5000+).

	Flavor	OS	CPU	OK	WARN	ERROR	Total
1	r-devel	Linux	ix86	769	135	32	936
2	r-devel	Linux	x86_64	523	101	36	660
3	r-prerel	Linux	ix86	816	154	30	1000
4	r-prerel	MacOS_X		ix86	738	160	95
5	r-prerel	Windows		x86_64	787	142	36
							993
							965

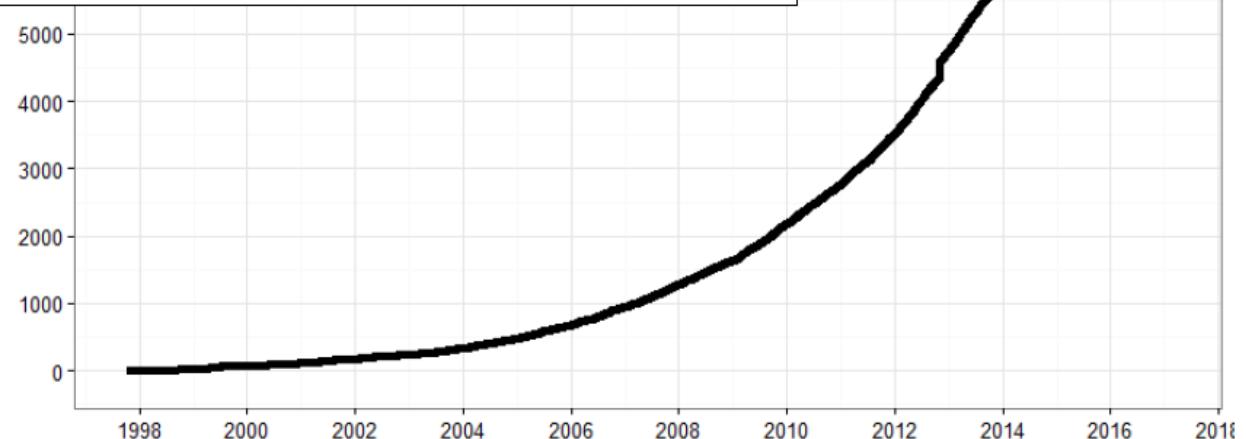


Number of R packages ever published on CRAN

CRAN now has 10,000 R packages. Here's how to find the ones you need.

January 27, 2017

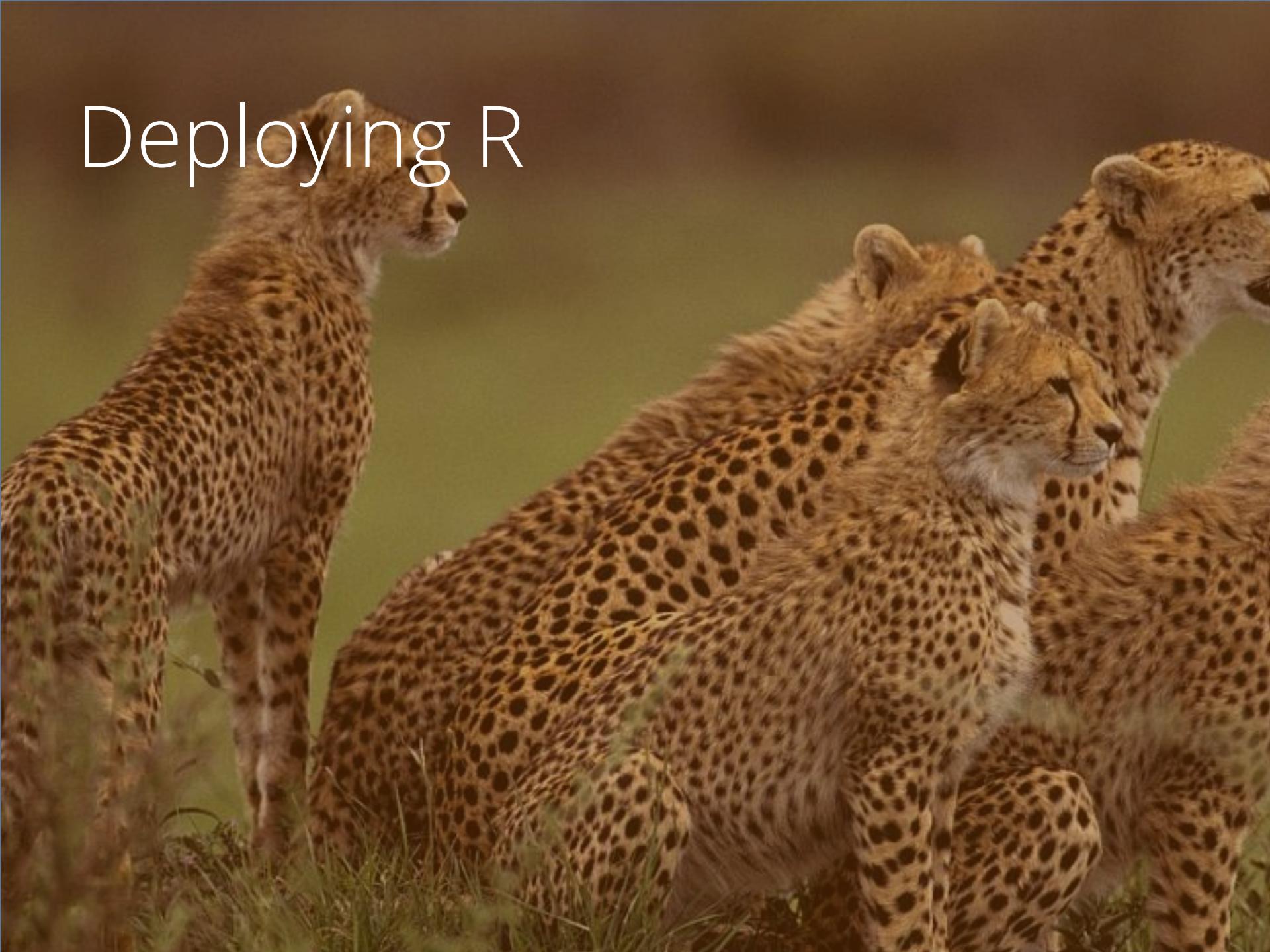
By David Smith



Some R Users



Deploying R

A photograph of two cheetahs in a grassy field. One cheetah is in the foreground, looking towards the right, while the other is partially visible behind it, looking towards the left.

Methods of Deploying R

- Scripts (batch mode)
- Packages
- Shiny apps
- APIs
- In DB



What is Shiny?

- R Package for interactive web apps developed by RStudio
- Gives the power of R in a convenient user interface
- Can be written entirely in R



A Basic Shiny App

- A basic app requires:
 - A user interface script
 - A "Server" script
- Runs using the `runApp` function



Shiny Case Study 1 – What and Why?

- The customer (?)
- A few hundred analysts
- SAS dependent
- New platform – decision to migrate had already been made



The Project

■ Application Development

■ Training users

■ Support

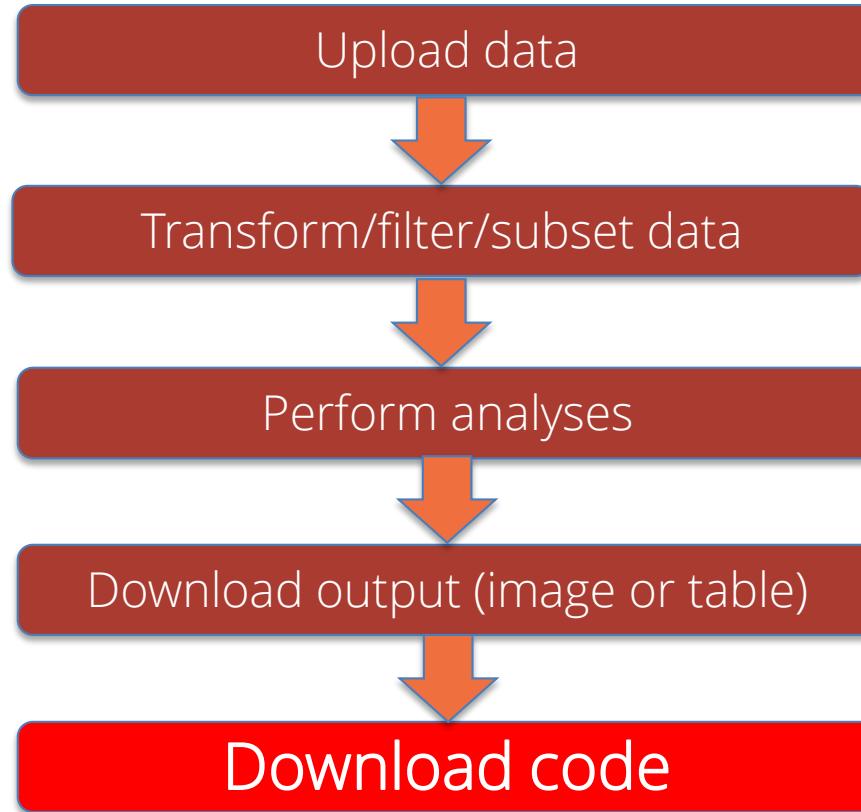


The Challenge

- User expectation
- Requirements
- Producing a learning tool



Workflow



The Application

Wafer Maps

Analysis/Plot Type
Chip site wafer map

Comparison Statistic

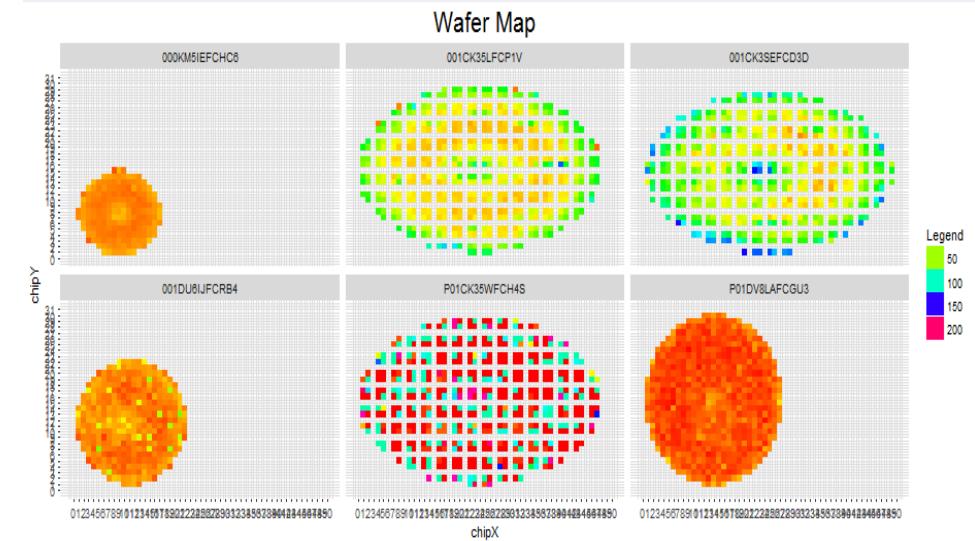
Min dimensions 0 **Max dimension** 10 **X axis** chipX

by 0 **by** 10 **Y axis** chipY

Midpoints
Auto-scale

Select a colour scheme
Default

Merge data
 No file selected



The Solution

Application



R script



Result

- Initially?
 - Continuity of business
- Culture Change
- Recruitment



Shiny Case Study 2 – What and Why?

- Pfizer Clinical Pharmacology group
- Based in East Coast US
- Wanted to be able to compare data from multiple clinical trials in a single application



Data Clean and View



Data Import and Export

Summary Statistics

Plot Viewer

Longitudinal Viewer

Forest Viewer

Endpoint:

ADAS-Cog/11

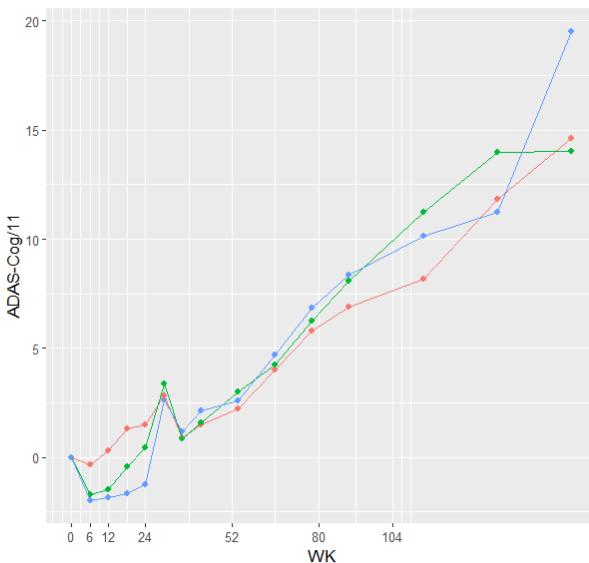
Plot options



Plot help



120: Burns A 2007



Related studies

237: Burns A 1999

Show 10 entries

Search:

DSID	DSIL	PY	SA	DOIURL	STD.IND	STD.DISCLASS
1	120	237	2007	http://...	Alzhei...	Mild-M...
2	120	237	2007	http://...	Alzhei...	Mild-M...
3	120	237	2007	http://...	Alzhei...	Mild-M...
4	120	237	2007	http://...	Alzhei...	Mild-M...
5	120	237	2007	http://...	Alzhei...	Mild-M...
6	120	237	2007	http://...	Alzhei...	Mild-M...
7	120	237	2007	http://...	Alzhei...	Mild-M...



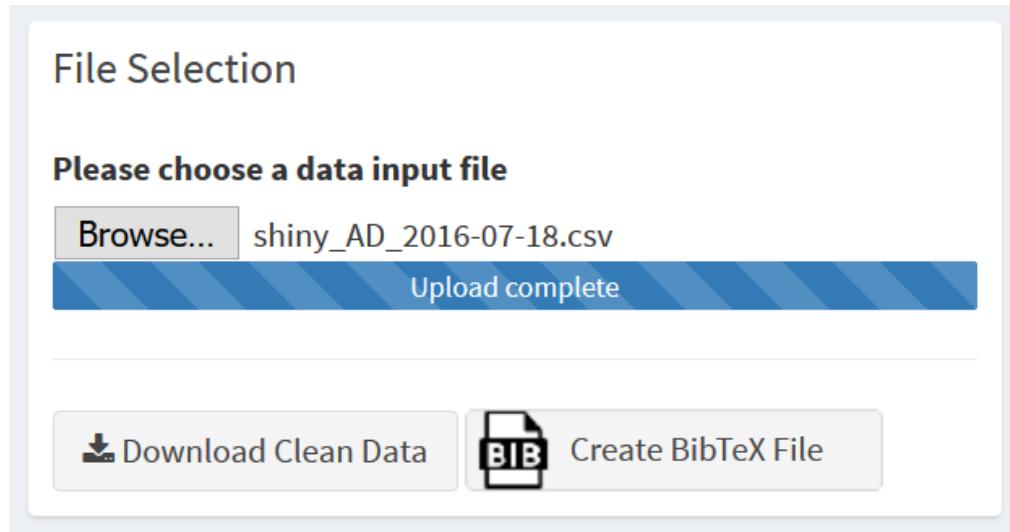
My favourite feature

File Selection

Please choose a data input file

[Browse...](#) shiny_AD_2016-07-18.csv
Upload complete

[Download Clean Data](#) [Create BibTeX File](#)



No more
tedious
bibliography
creation

One Click Bibliography



Result

- Time saving
 - Everything in one place makes cleaning and analysis more efficient
 - Much quicker to navigate between the paper and the data and all papers
- More efficient decision making



What is an API?

- “a set of clearly defined methods of communication between various software components”
- Allows access to something whilst governing things like usage rate/content



APIs in R: Plumber



Plumber



```
addVals.R *  
Source on Save |        
```

```
1 #* @get /sum  
2 addVals <- function(a, b){  
3   as.numeric(a) + as.numeric(b)  
4 }  
5
```

```
R plumb.R x [ ] Source on Save |     
1 library(plumber)  
2 r <- plumb("addVals.R")  
3 r$run(port=8000)  
4
```

```
$ curl "http://localhost:8000/sum?a=1?b=2"  
[3]
```

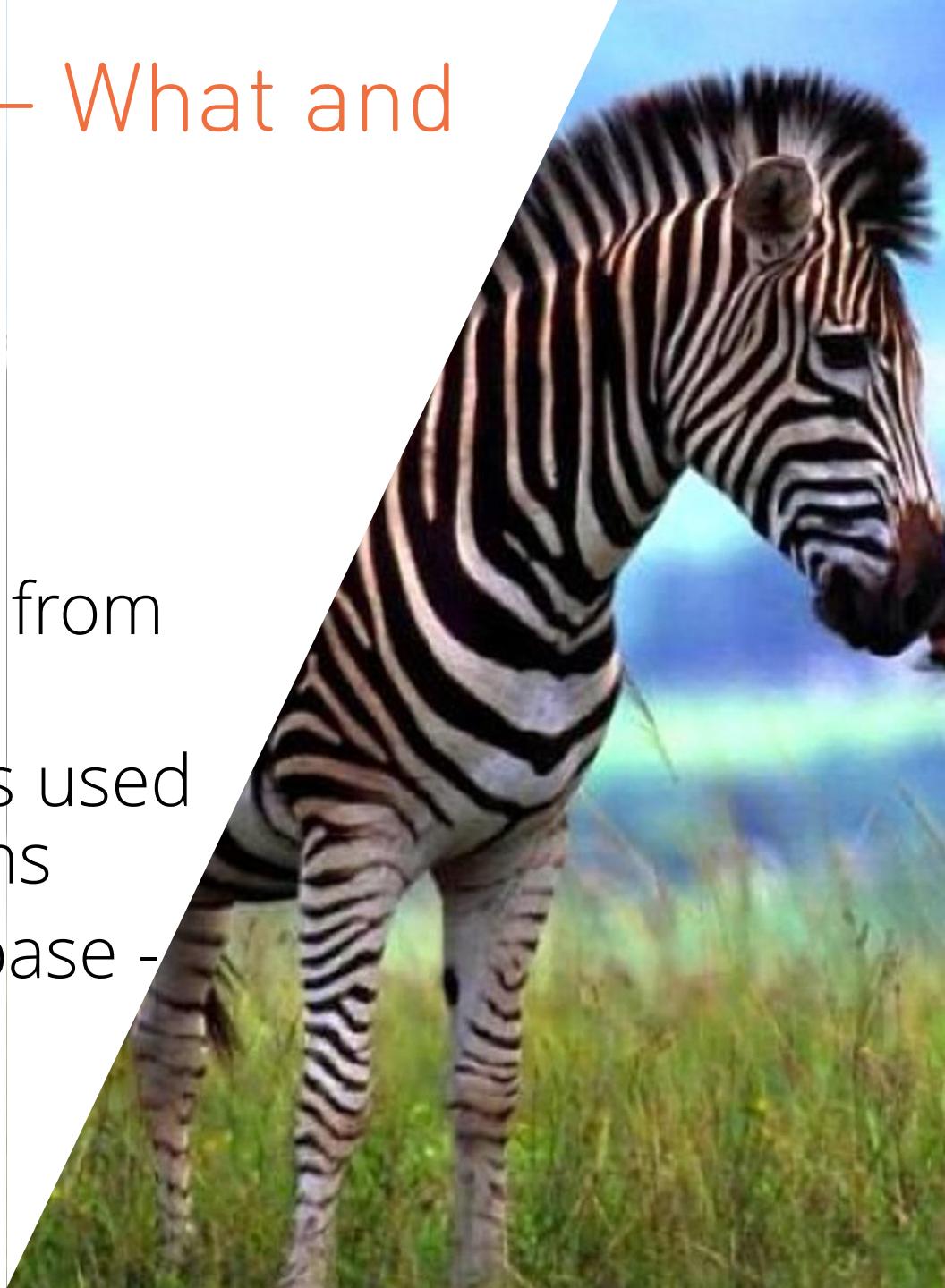


API Case Study – What and Why?

Office for National Statistics (ONS)

Desire to move away from SAS, but:

- Mix of technologies used in some applications
- Large legacy codebase - where to start?



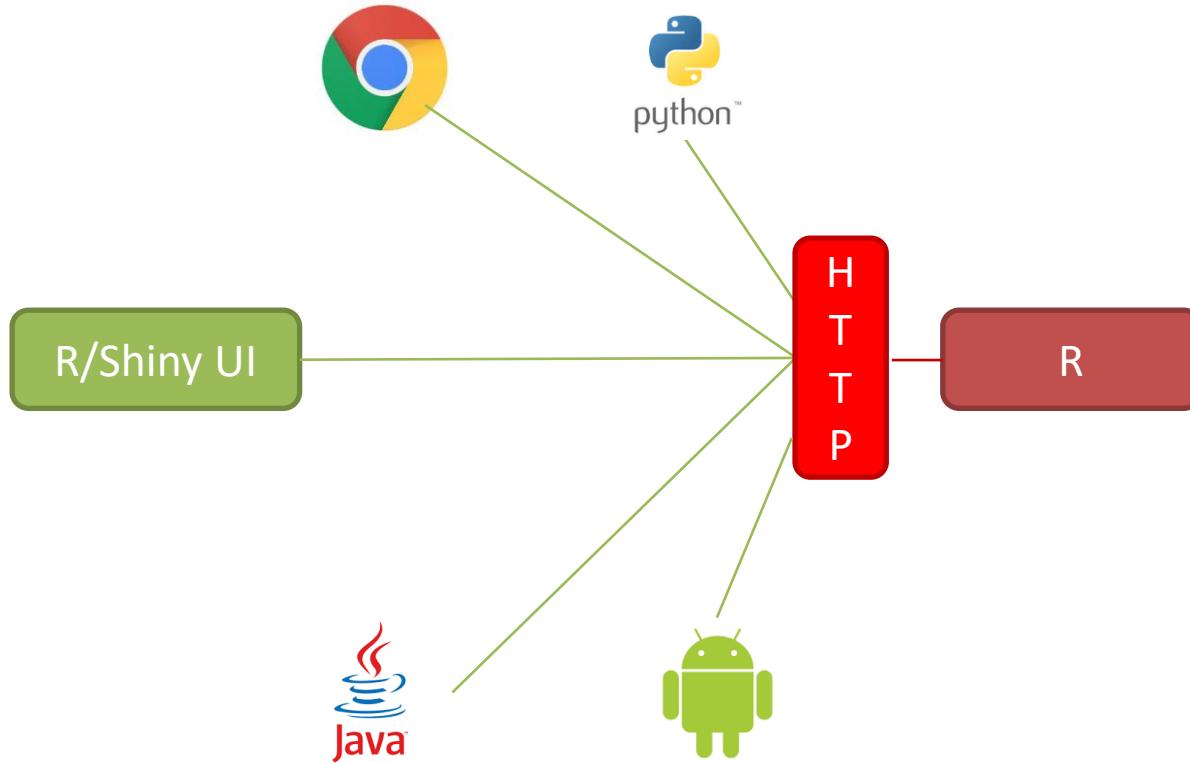
The Solutions

1. Proof of concepts
 - Microservices (API)

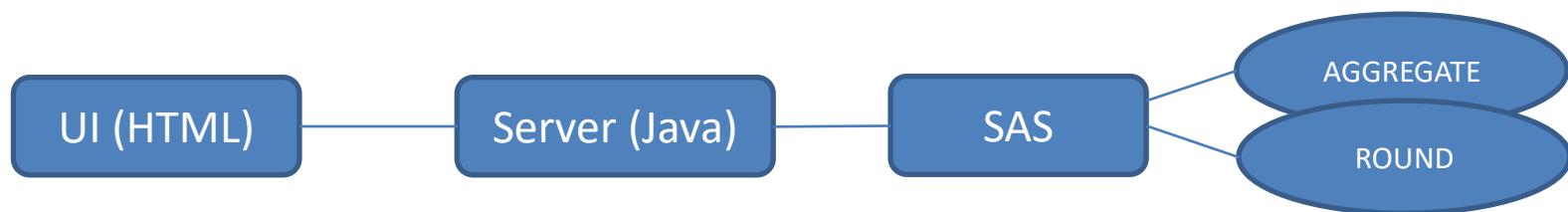
2. Analysis of existing systems
 - sasMap



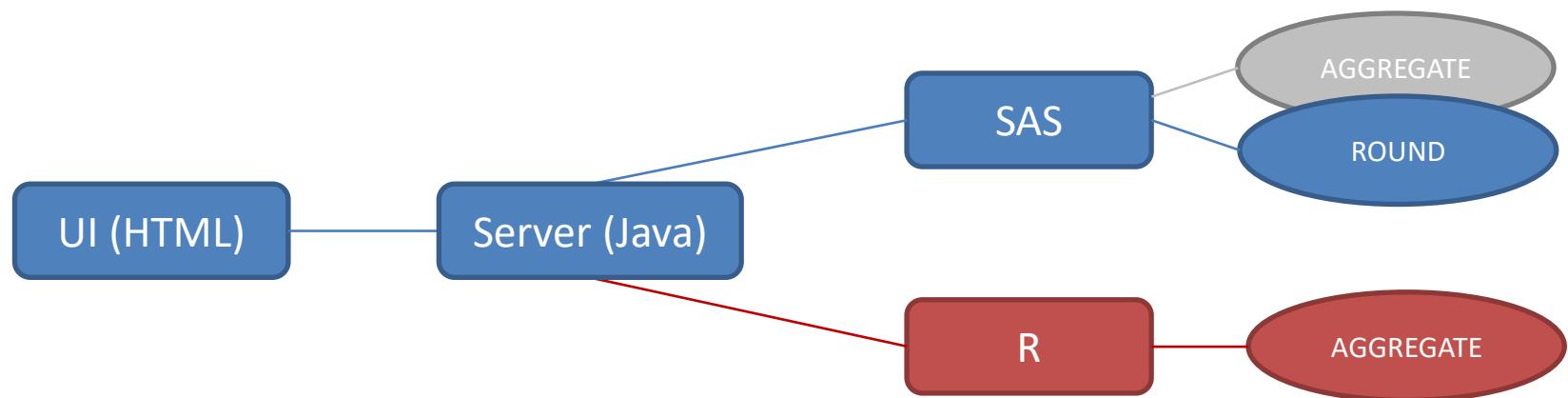
Solution 1: Microservices



Existing Architecture



PoC Architecture

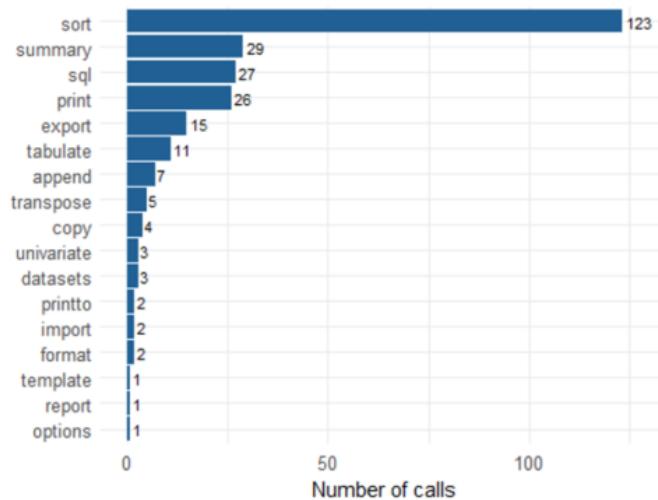


Solution 2: sasMap

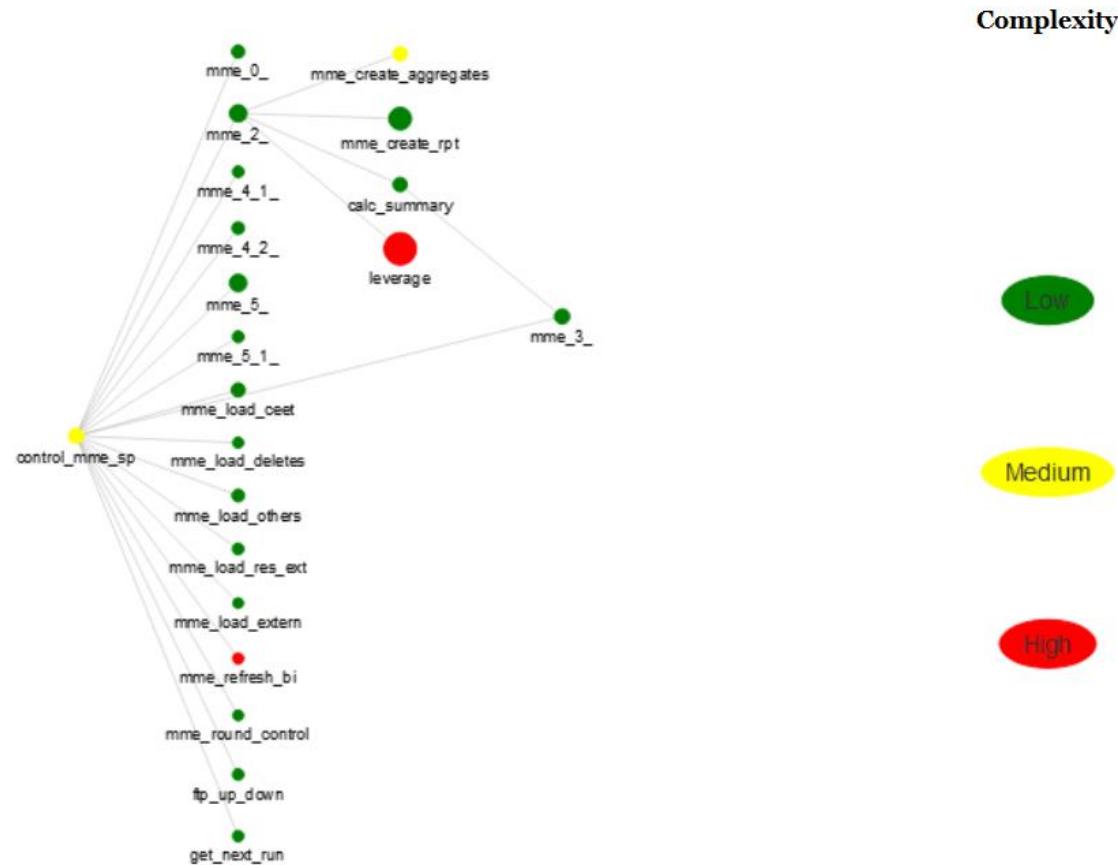
Procedure Calls

Metric	Value
Proc calls	262
Unique proc calls	17

The plot below shows the most common procedure calls in the application.



Solution 2: sasMap



Result

- Proof of concept has been handed over and ONS now looking at expanding and implementing
- SAS code analysis docs being used to guide decision on prioritising and resourcing for replacement of other applications



Overcoming Misconceptions



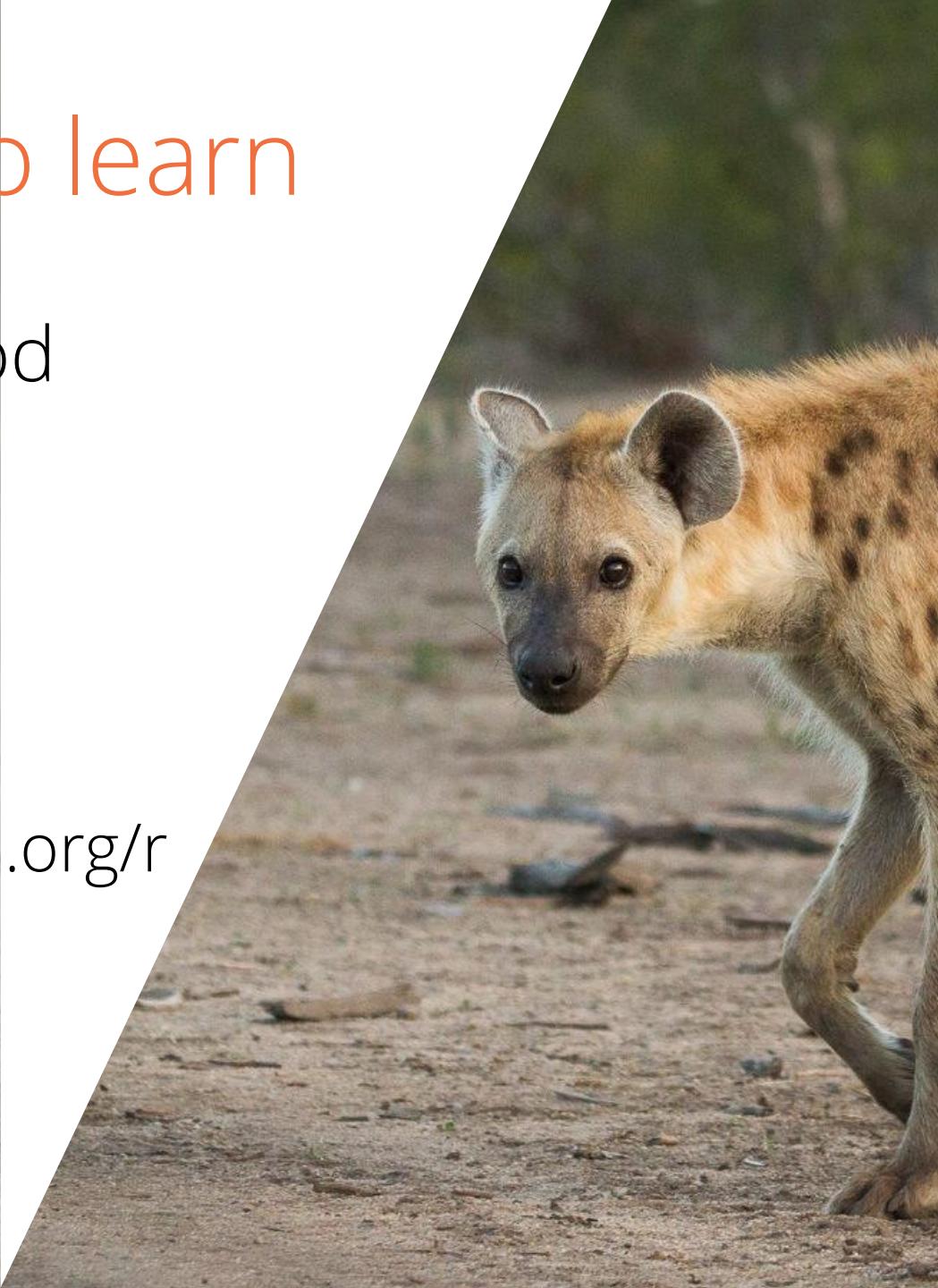
#1: You can't trust open source

- Mindset shift
- ValidR
- All software has disclaimers (e.g. SAS)



#2: R is hard to learn

- Depends how good your teacher is! ;)
- Tidyverse
 - See varianceexplained.org/r/teach-tidyverse/



#2: R is hard to learn

```
> airquality [ airquality$Wind > 15, ]  
> filter(airquality, Wind > 15)      # dplyr  
> airquality %>% filter(Wind > 15)  # magrittr  
  
> x <- 1:10  
> x = 1:10
```



#3 But there's no tech support!

- Mindset shift again
- R user community
- <https://community.rstudio.com/>



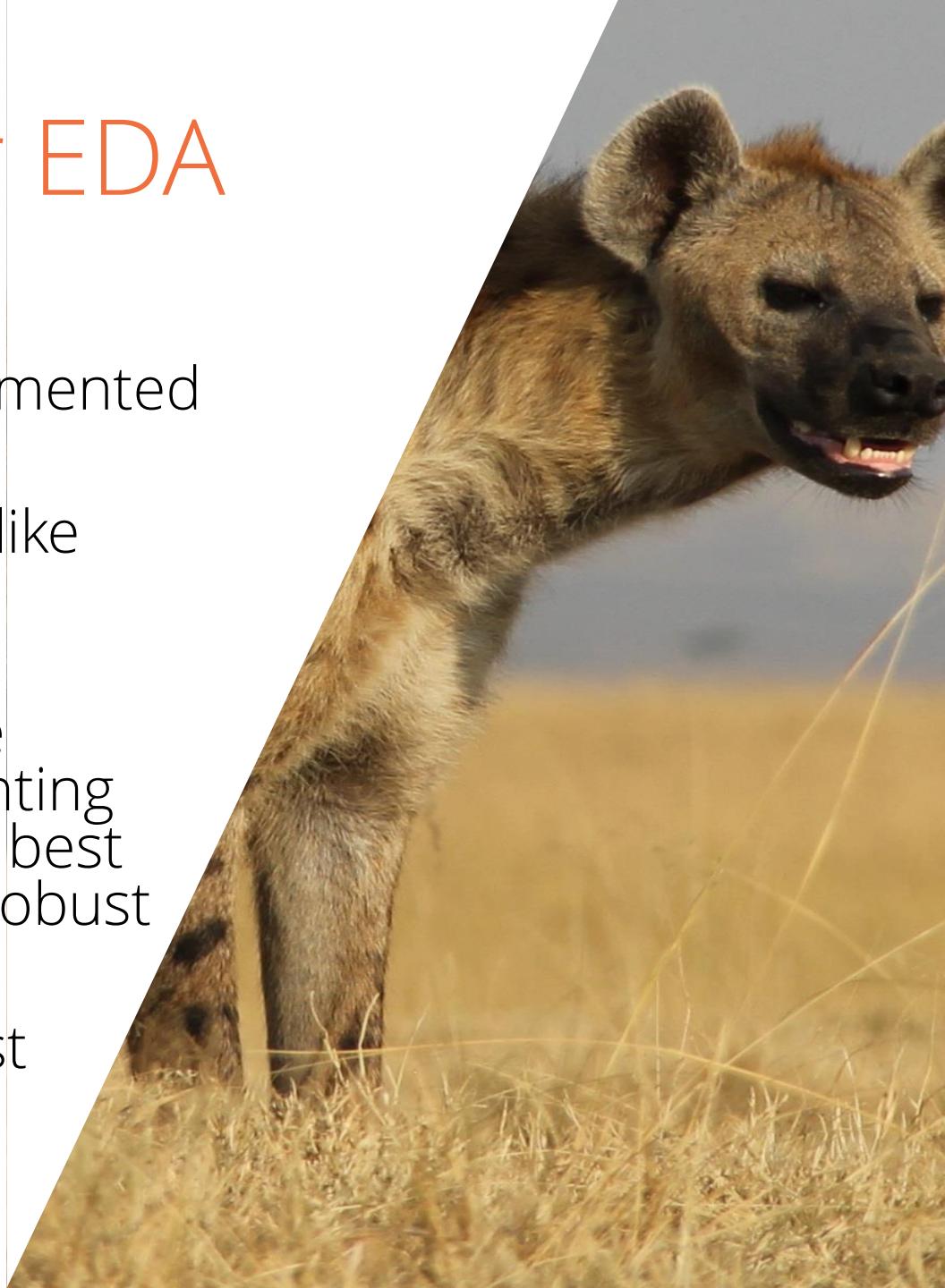
#4 But my results will be different!

- Sometimes true!
- 0.5 rounding vs SAS
- But typically the same, or better methodologies used, and can test/compare.



#5 R is just for EDA

- Good at this, but rich modelling suites implemented
- Connections to things like Tensorflow, Spark etc
- testthat and R package structure for implementing software development best practices and making robust
- Scale without extra cost



#6 R is limited by memory/is slow

- Can be if you write for loops within for loops...please don't!
- Data.table, sparklyr
- Change in mindset – in DB computing
- Speed – faster than used to be



#7 Python is more popular so it's better to use that

Language Rank	Types	Spectrum Ranking
1. C	📱💻⚙️	100.0
2. Java	🌐📱💻	98.1
3. Python	🌐💻	98.0
4. C++	📱💻⚙️	95.9
5. R	💻	87.9
6. C#	🌐📱💻	86.7
7. PHP	🌐	82.8
8. JavaScript	🌐📱	82.2
9. Ruby	🌐💻	74.5
10. Go	🌐💻	71.9



Thanks!



@nic_crane



ncrane@mango-solutions.com

