

2<sup>nd</sup> Aug '16

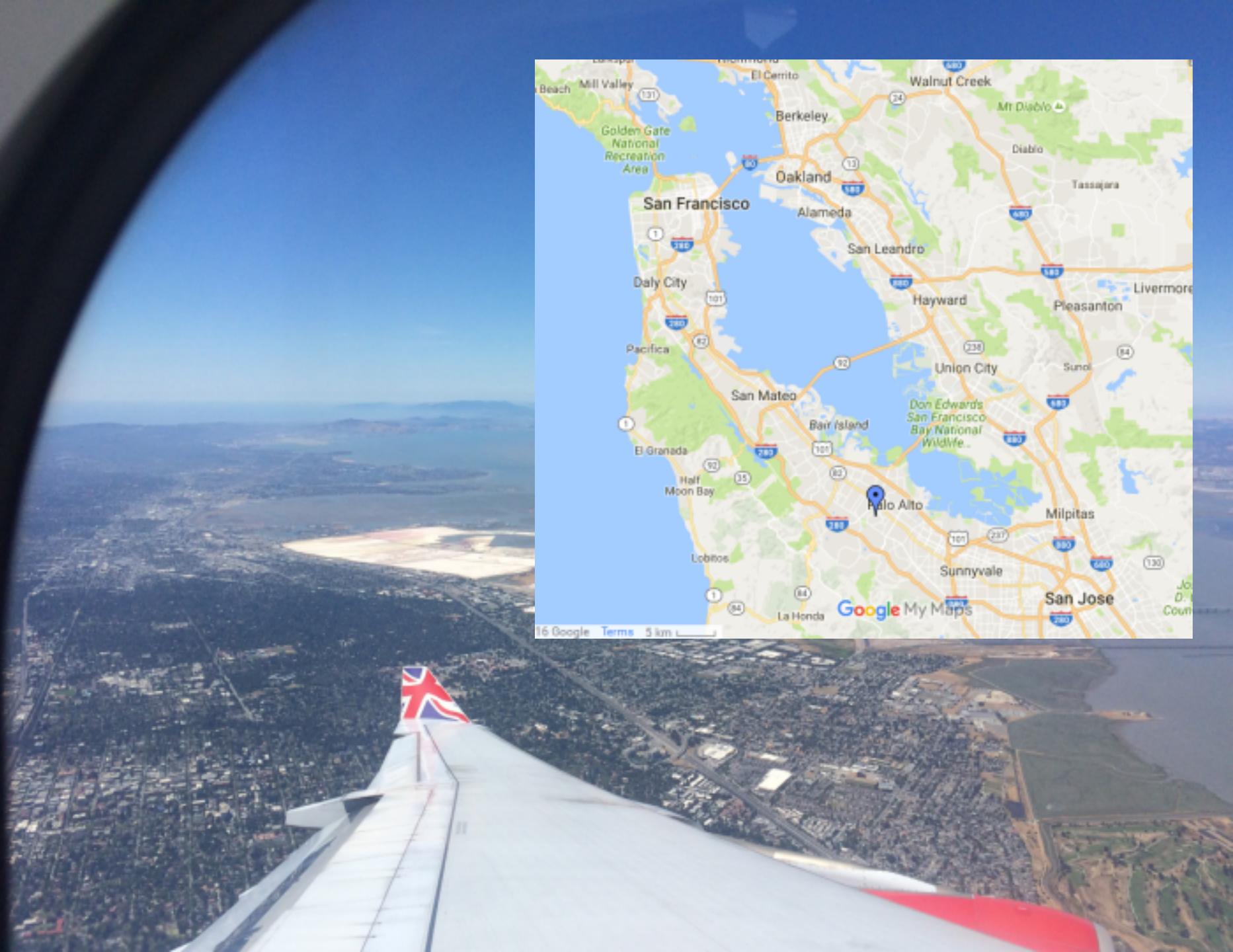


## Review

Alice Daish & Hannah Frick



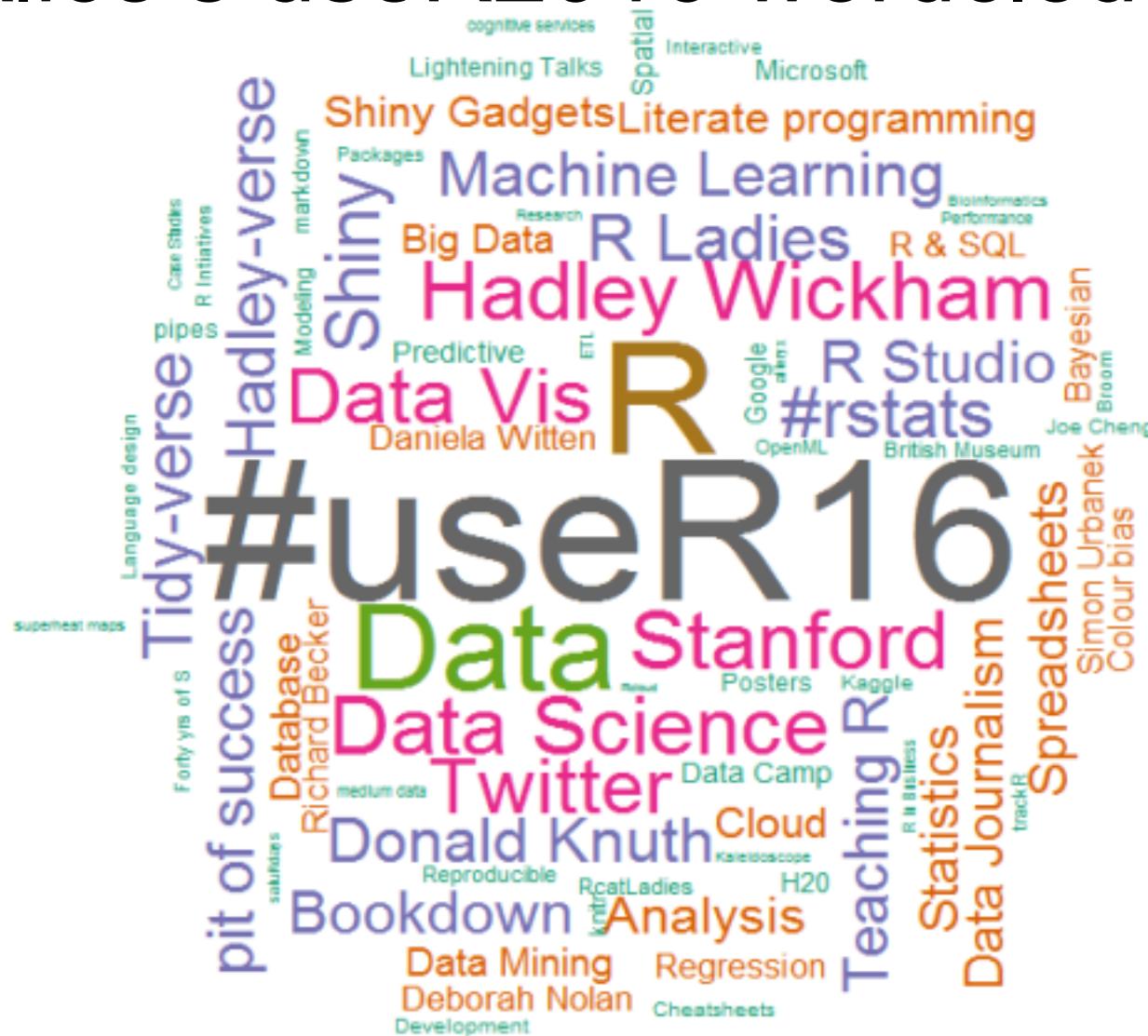
-Ladies London





reproducible research  
Bioconductor algorithm  
randomization interactive graphics unsupervised learning  
analytics XML natural language processing  
parallel processing data science  
rjson psychometrics experimental design  
CRAN education k-means Hadoop  
genomics networks user random forest  
R-markdown high-performance computing  
Shiny R-cloud survey dendrogram big data LASSO  
ETL high-dimensional data science  
Spark GUI vignette adaptive learning heat maps  
permutation tests Bayes Rcpp bioinformatics json rgl  
Python visualization dashboards  
optimization asynchronous hackathon  
simulation ensemble learning neural networks  
ggplot2 text mining penalized likelihood  
automation statistical inference forecasting  
generalized additive models digital humanities grid computing  
machine learning OpenStreetMap

# Alice's useR2016 wordcloud :)



# useR! 2016 Program at a glance

	Monday, June 27, 2016	Tuesday, June 28, 2016	Wednesday, June 29, 2016	Thursday, June 30, 2016
8:00-8:30		Registration (all day)	Registration (morning)	
8:30-9:00	Registration (all day)	Opening Session		
9:00-10:00		Richard Becker	Hadley Wickham	Deborah Nolan
10:00-10:30		Coffee Break including Poster Exhibits (display at Sponsor Pavilion)		
10:30-11:00	Morning Tutorials (including coffee break)	<ul style="list-style-type: none"> <li>Bayesian (Barnes)</li> <li>R in Business (Econ 140)</li> <li>Bioinformatics (Lane)</li> <li>Kaleidoscope (McCaw)</li> <li>Performance (Siepr 130)</li> </ul>	<ul style="list-style-type: none"> <li>Statistics Methods (Barnes)</li> <li>Performance (Econ 140)</li> <li>Bioinformatics (Lane)</li> <li>Kaleidoscope (McCaw)</li> <li>Lightning Talks (Siepr 120)</li> <li>Database (Siepr 130)</li> </ul>	<ul style="list-style-type: none"> <li>Teaching (Barnes)</li> <li>Lightning Talks (Econ 140)</li> <li>Statistics &amp; Big Data (Lane)</li> <li>Kaleidoscope (McCaw)</li> <li>Graphics (Siepr 130)</li> </ul>
11:00-12:00				
12:00-1:00	Lunch	Lunch including Poster Exhibits (displayed at Sponsor Pavilion)		
1:00-2:00		<ul style="list-style-type: none"> <li>Regression (Barnes)</li> <li>R &amp; Other Languages (Lane)</li> <li>Kaleidoscope (McCaw)</li> <li>Case Study (Siepr 120)</li> <li>Teaching (Siepr 130)</li> </ul>	<ul style="list-style-type: none"> <li>Reproducible Research (Barnes)</li> <li>Generalized Mixed Models (Econ 140)</li> <li>Sponsor Session (Lane)</li> <li>Kaleidoscope (McCaw)</li> <li>Spatial (Siepr 120)</li> <li>Packages &amp; Development (Siepr 130)</li> </ul>	<ul style="list-style-type: none"> <li>Sponsor (Lane)</li> <li>Lightning Talks (McCaw)</li> <li>Lightning Talks (Siepr 130)</li> </ul>
2:00-2:15	Afternoon Tutorials (including coffee break)			
2:15-2:30				Simon Urbanek
2:30-3:00		Coffee Break including Poster Session (displayed at Sponsor Pavilion)		
3:00-3:30				Closing Remarks
3:30-3:45				
4:00-4:15	Short Break	Donald Knuth	Daniela Witten	
4:15-4:30				
4:30-4:45		Short Break		
4:45-5:30	R Initiatives	<ul style="list-style-type: none"> <li>Analytics (Barnes)</li> <li>Miscellaneous (Econ 140)</li> <li>Interactive (Lane)</li> <li>Kaleidoscope (McCaw)</li> <li>Miscellaneous (Siepr 120)</li> <li>Machine Learning (Siepr 130)</li> </ul>	Bus to Cruise (board by 5:00 latest)	Conference ends
6:00-6:15			Cruise and Conference Dinner (Sponsored by Microsoft)	
6:30-8:30		Welcome Reception (Sponsored by RStudio)	Return to campus at 10:30	



# Machine Learning Tutorial

## Erin LeDell @ledell (R-Ladies SF)

six popular supervised machine learning methods:

- Classification and Regression Trees (CART)
- Random Forests (RF)
- Gradient Boosting Machines (GBM)
- Generalized Linear Models (GLM)
- Deep Neural Networks (DNN)
- Stacking / Super Learner (SL)

Here are some practical, related topics we will cover for each algorithm:

- Dimensionality Issues
- Sparsity
- Normalization
- Categorical Data
- Missing Data
- Class Imbalance
- Overfitting
- Software
- Scalability

<https://github.com/ledell/useR-machine-learning-tutorial>

# Other Tutorials

- Using Git and GitHub with R, RStudio, and RMarkdown - Jennifer Bryan
- Handling and Analyzing Spatial, Spatiotemporal and Movement Data - Edzer Pebesma
- Dynamic Documents with R Markdown - Yihui Xie

# Keynotes

- Forty Years of S - Richard Becker
- Literate Programming - Donald Knuth
- Statistical Thinking in a Data Science Course  
- Deborah Nolan
- Tidy tools for data science - Hadley Wickham
- Flexible and Interpretable Regression Using  
Convex Penalties - Daniela Witten
- RCloud - Simon Urbanek

# Forty years of S - Richard Becker

What was coding like back then? Well batch coding was punch cards and print outs #useR2016

S(R) Language design history : simple formal grammar, it remembers data types, powerful subscripting #useR2016

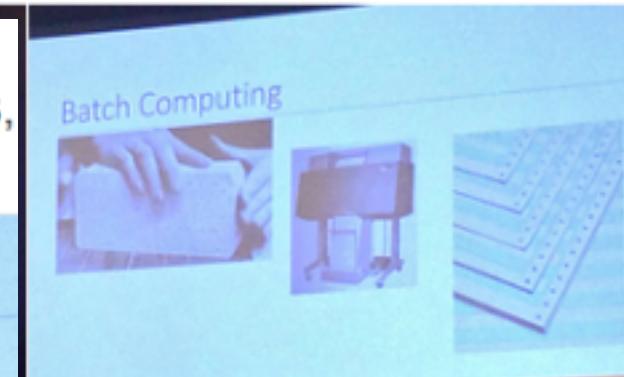
Language Design

- Simple formal grammar
- No declarations
- Primitive data types: numeric, character string, logical
- Composite data: lists with named components
- Powerful subscripting
- Persistent storage
- No artificial constraints

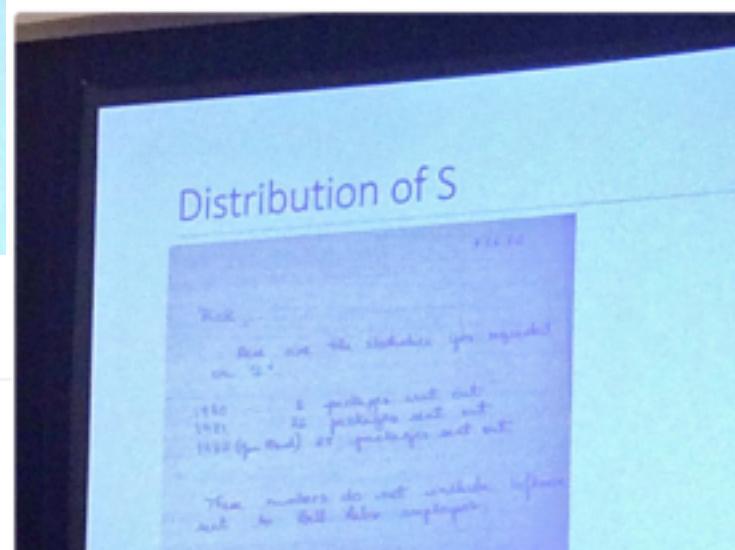
Hadley Wickham @hadleywickham · Jun 28  
2 of the 5 original authors of S were women #rstats #user2016

14 109 151 ...

Alice Data @alice\_data · Jun 28  
Why has S survived 40yrs? "We built it for us, no junk code, re-written 3times, and then along came R" #useR2016



The classic S download "letter of tape mail outs 1980 - 8, 1981 - 26, 1982 - 25" #useR2016



# Literate Programming - Donald Knuth



# Interactive graphics - Hadley Wickham

@hadleywickham keynote at #useR2016

"my goal is to make a pit of success" by  
tidy data and tidy tools #DataScience



# Stats Thinking in Data Science Course

- Deborah Nolan



# Stats Thinking in Data Science Course

## - Deborah Nolan

Just imagine #stats course without simplified scenarios, canned data, non-coding, & always normal distrib  
#useR2016



Teach #stats in #DataScience update your practises & here R some suggestions & examples via Deborah Nolan #useR2016

NEED: Update Instructional Practice

In the beginning we taught mathematics and called it statistics; much of this was probability. Then with the help of computers, we started to teach data analysis and statistical modelling; this was fine apart from one feature: it was largely context free. (Speed, 1986)

traditional statistics courses do “not attempt to teach what we do, and certainly not why we do it... these courses are caught in a time warp that bores teachers and subsequently bores students.” (Efron, 2003)

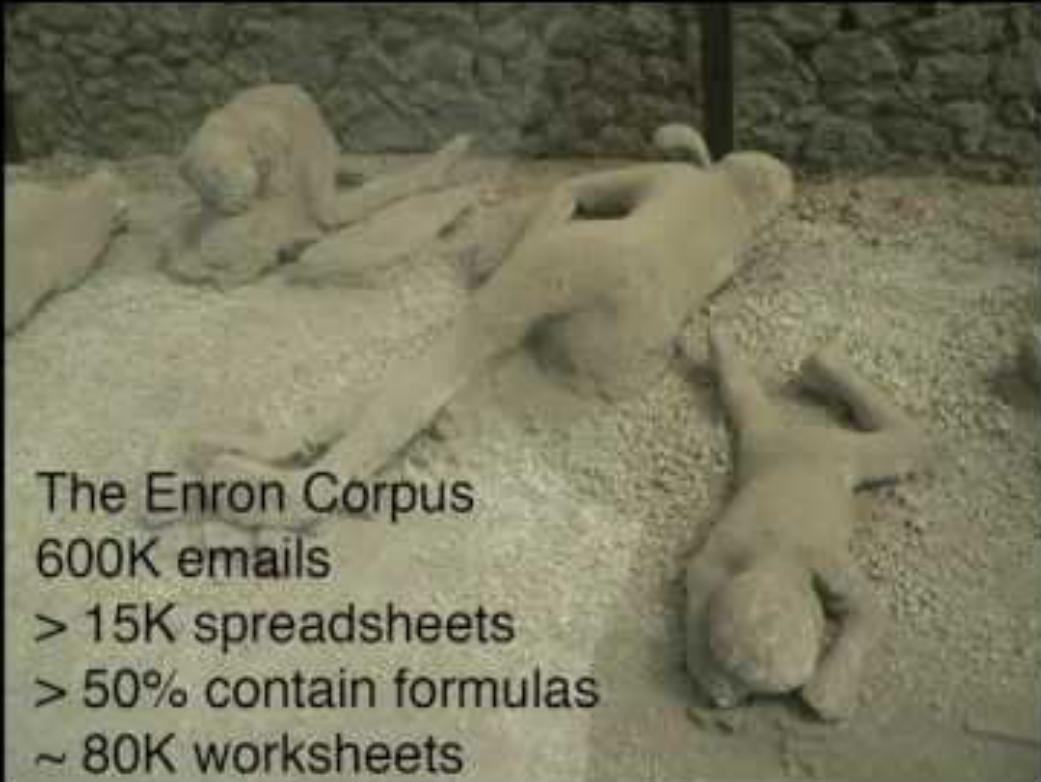
Suggestions for Updating Instructional Practice

- Provide Open-ended Problems with Context
- Experience Complete Data Analysis Process including acquisition, preparation, transformation of data
- Conduct Modern EDA throughout Data Analysis Process
- Emphasize Core Computational Paradigms
- Reframe Statistical Theory from Normal Theoretical Perspective

A Few Concrete Examples

- Full Case Study with data access, cleaning and transformation
- Instruction-Student Cycle of Exploration
- Classroom programming activities
- Replace and Compose Visualization exercises
- Integrate the 3Rs into theoretical statistics

# Spreadsheets - Jenny Bryan



The Enron Corpus  
600K emails  
> 15K spreadsheets  
> 50% contain formulas  
~ 80K worksheets

# Alice Fav Presentation



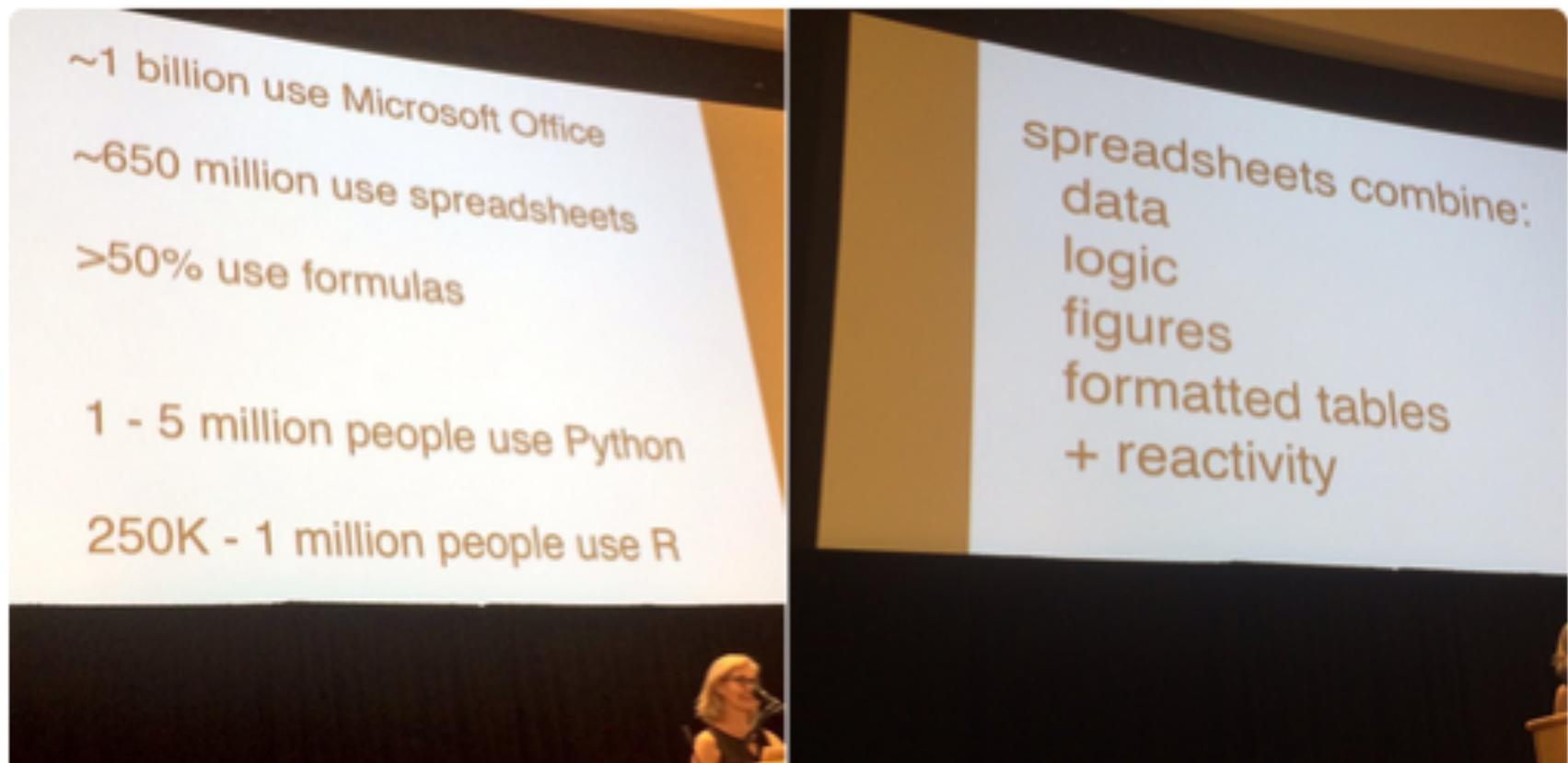
**Alice Data** @alice\_data - Jun 28

Ahh spreadsheets talk from @JennyBryan @STAT545 love it "some of my best friends use spreadsheets" #useR2016



# Alice Fav Presentation

#rstats are drastically outnumbered by ~650m using spreadsheets- why?  
bc it can combine data & is reactive #useR2016



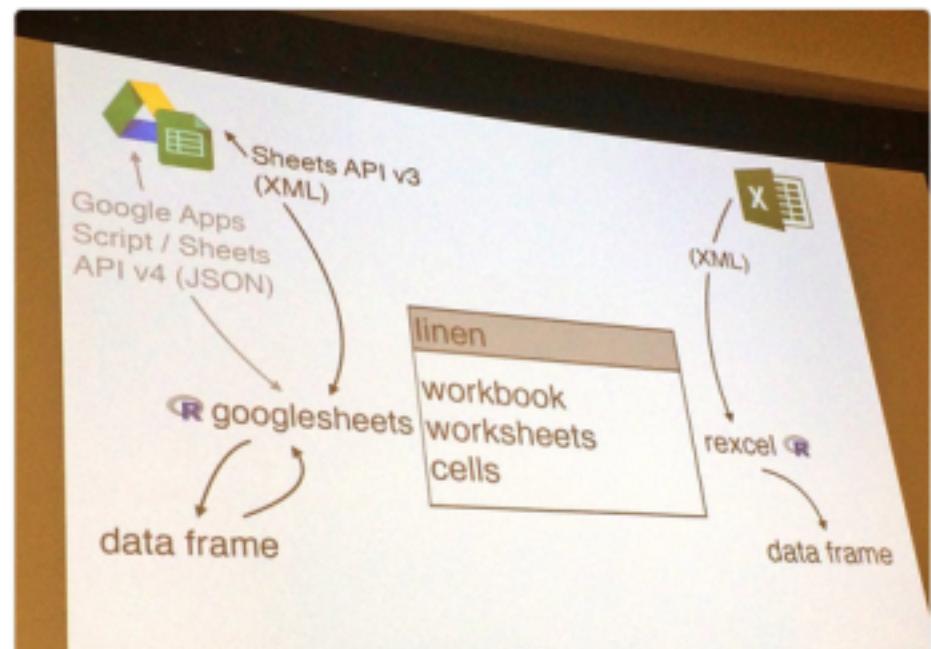
# Alice Fav Presentation

Can't wait  
for @rsheets  
to be a reality



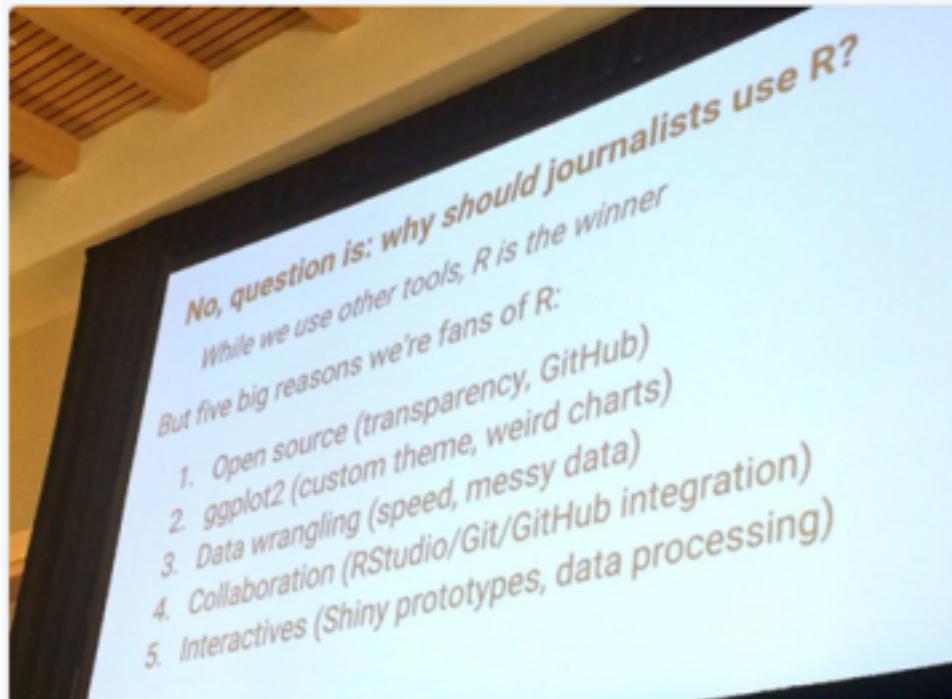
Alice Data @alice\_data · Jun 28

Hours that could be saved using future  
`#linen` package wit formatted  
spreadsheets capturing behaviour  
`#userR2016`

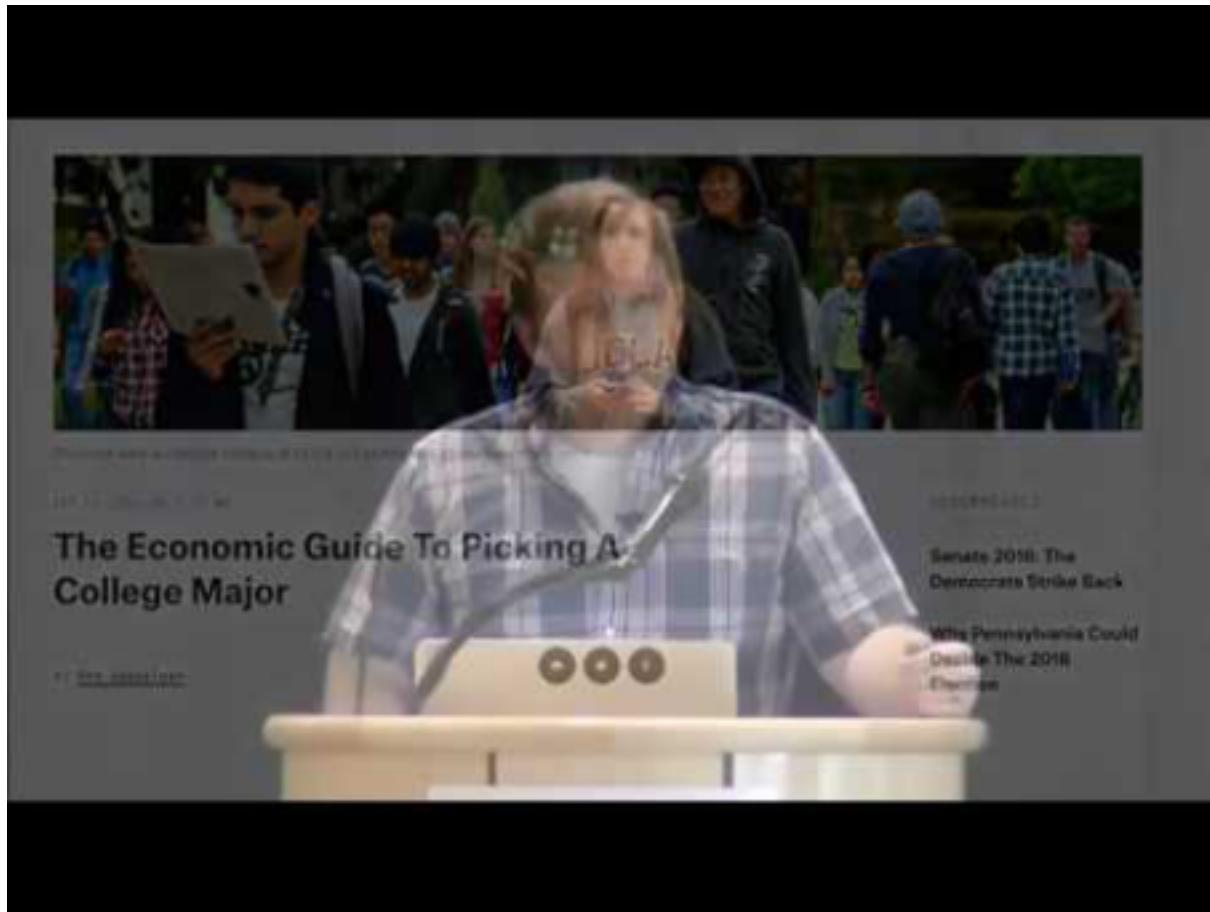


# Data Journalism

Why do #datajournalist use #rstats?  
#useR2016

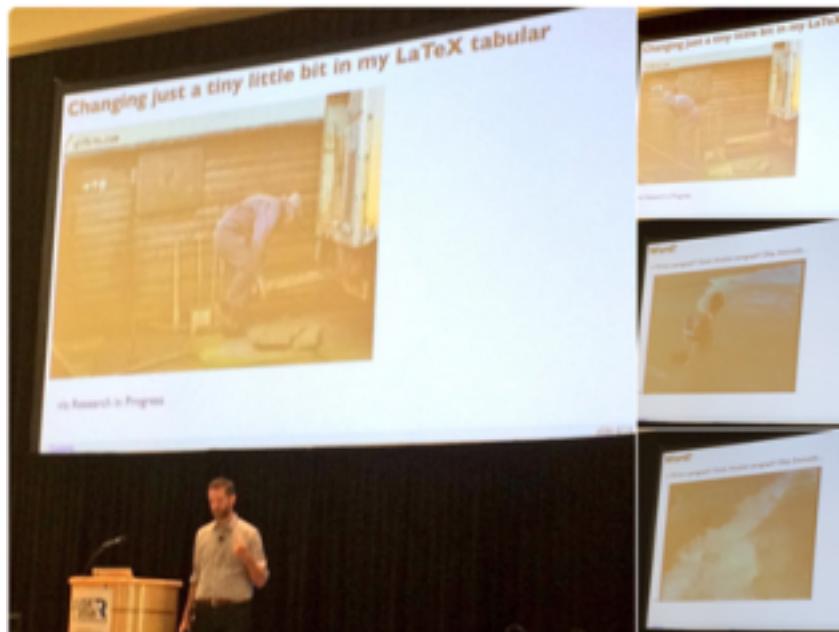


# Data Journalism



# Bookdown

Excellent @hadleywickham karaoking  
@xieyihui gif slides on authoring using  
#rmarkdown #useR2016



Check out #bookdown for writing books in  
#Rmarkdown bookdown.org #useR2016  
#rstats

# Careful with colours

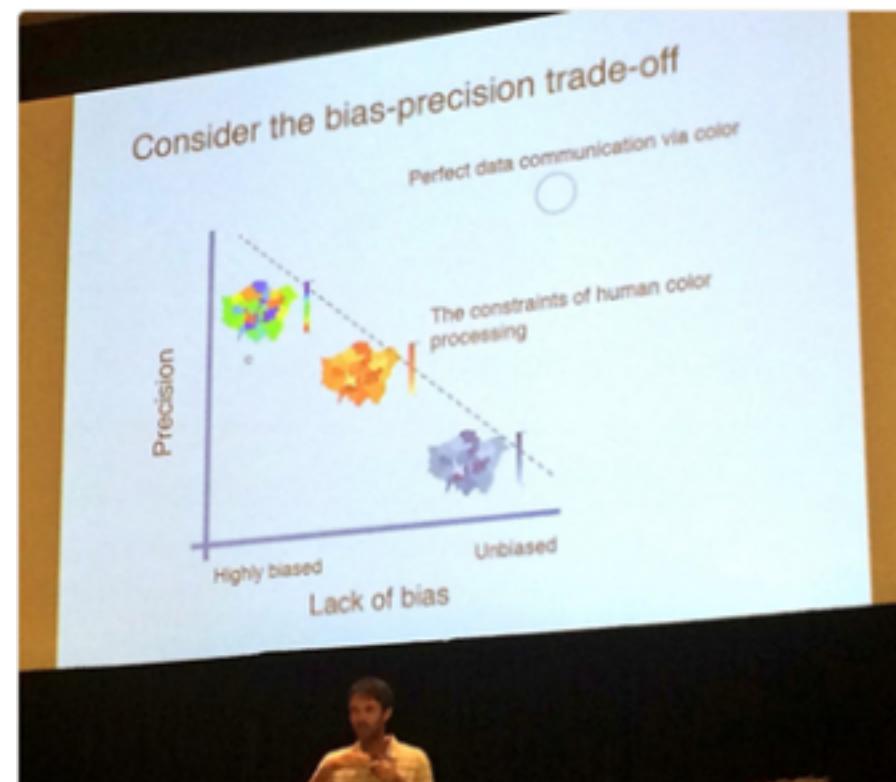
Check your colours! If you're worried about bias & precision interpretation of colours exists! #useR2016



@will\_cornwell

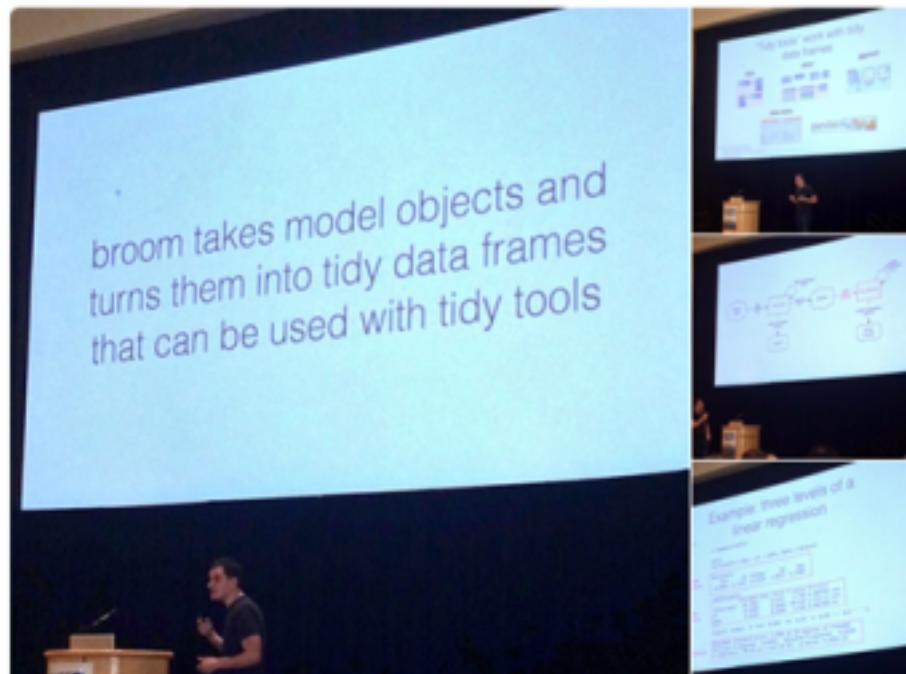
Wow learnt something new Russians have 2 types of Blue (light & dark) compare to others who only had one "blue" #dataviz #fact #useR2016

Think about the bias-precision trade-off  
#rstats #dataviz #useR2016



# Broom

Love **#broom** ! I need to do some tidying I  
make a lot of mess! messy hacker!  
**#useR2016**



# RCloud

Final #useR2016 keynote from Simon Urbanek "RCloud" - #rstats

anywhere 🌎 anytime ⏱ via the cloud ☁️ !

The image shows a presentation slide titled "RCloud" on the right and a video feed of a speaker on stage on the left. The slide has a blue header with the title "RCloud" and a subtitle "Collaborative Environment for Visualization and Big Data Analytics". It features a bulleted list of benefits:

- Push data and analytics “in the cloud”
  - user doesn’t care where and how
  - both consumption and development in the could
  - solves scalability and reproducibility
- Everything is web-based
  - access from anywhere
  - leverage web-based visualizations and interactive graphics
- Sharing and collaboration
  - share both code and results
  - search, run, comment, fork, modify, ...

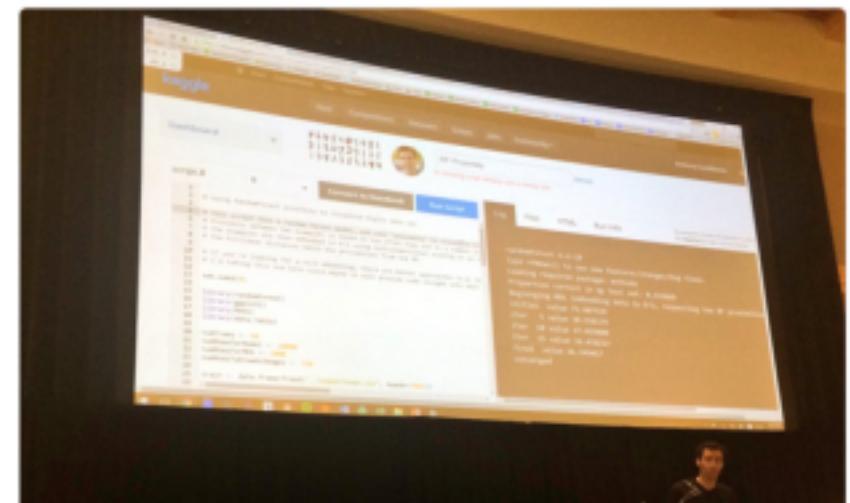
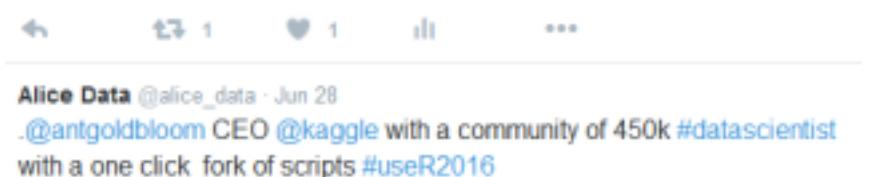
Below the slide, a video frame shows a man with a beard speaking at a podium on a stage, with an audience visible in the foreground.

# Try Kaggle

Learning about @kaggle 1000 machine learning comp per month and #rstats is the second most used scripts #useR2016

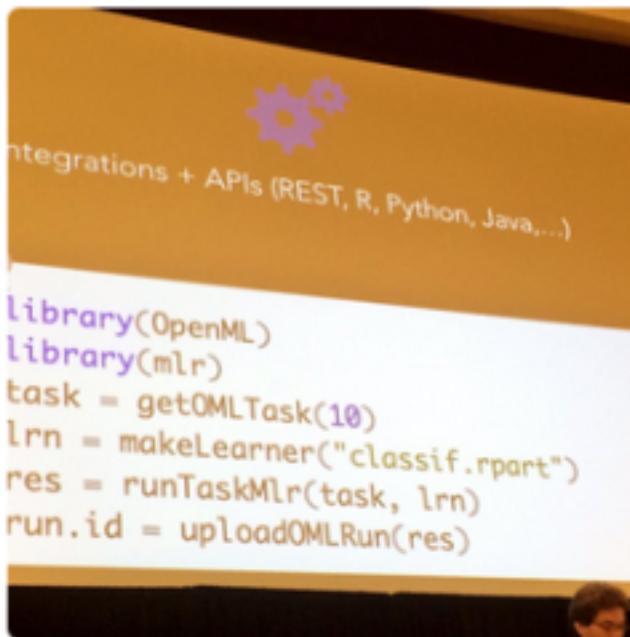


@kaggle supportive of dataset sharing, reproducibility & being a resource to see what other people have done with #rstats packages #useR2016



# OpenML

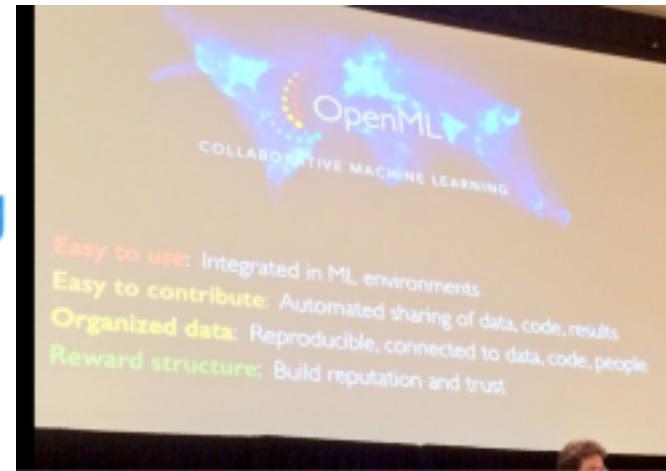
How to get started with open machine learning in #rstats #OpenML openml.org #useR2016



The OpenML website homepage. It features the OpenML logo and the tagline "Exploring machine learning better, together".

Key statistics displayed:

- 19592 data sets
- Find or add data to analyse
- 46702 tasks
- Download or create scientific tasks

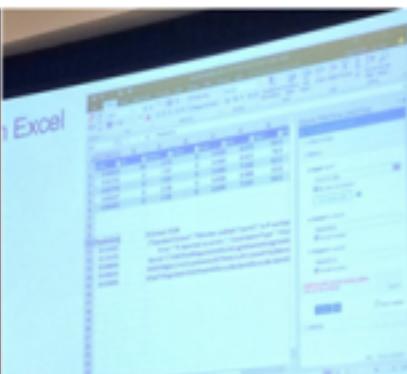
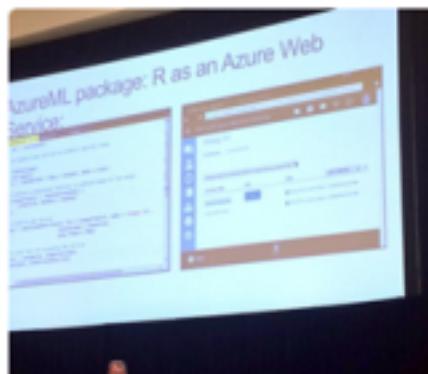


# R and Microsoft

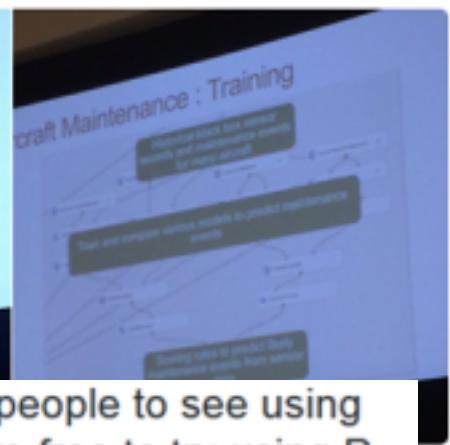
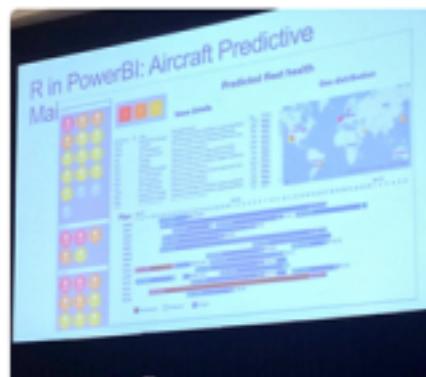
#rstats and @Microsoft building MicrosoftR making R fast @revodavid  
#useR2016



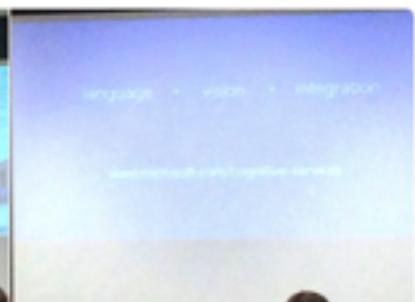
This is amazing from #rstats as web service and then doom into excel  
#useR2016



#powerBI driven by #rstats and then workflow creating R in Azure ML and turn it into API #useR2016



@revodavid helping people to see using #DataScience -API are free to try using R microsoft.com/cognitive-serv...  
#useR2016



# R-Consortium

What is @RConsortium ? joseph rickert @microsoft talking about the drive from business of #rstats growth #user2016

Goals of @RConsortium #useR2016



# Top areas of discussion

- Teaching R
- Tidy data / tidyverse
- Cloud
- R Consortium & Sponsors
- Machine Learning
- R-Ladies



# Data Camp

The screenshot shows the DataCamp homepage with a teal header. The header includes the DataCamp logo, navigation links for Home, Courses, Pricing, Business, Community, a Sign In button, and a prominent 'Create Free Account' button. Below the header, a large banner features the text 'THE EASIEST WAY TO Learn Data Science Online'. It includes a subtext about learning data analysis from the comfort of your browser at your own pace, tailored to your needs and expertise. Two buttons are visible: 'Start Learning R' (yellow) and 'Start Learning Python' (blue). To the right of the banner is a 'Create Your Free Account' form with fields for Email and Password, and a 'Get Started' button.

## Improve your skills - and your career

No matter what industry you're in, learning how to analyze and understand your data is critical. That's why DataCamp provides you with the tools to learn the data science skills you need to start your own data projects.

Check out @DataCamp excellent searchable interface for **#rstats** packages & teaching [rdocumentation.org](http://rdocumentation.org) **#useR2016**

The screenshot shows a search interface for DataCamp packages. The main heading reads 'and BioC packages.' Below it, statistics are provided: 9925 packages, 49070 package versions, and 16024 collaborators. A search bar contains the placeholder text 'Search for packages, fu' and a yellow 'Search' button. Below the search bar, text encourages exploring packages via 'Task Views.'

[user2016.org/](http://user2016.org/)

#useR2016review

#RLadiesLondon

# Resources

<https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016>

<http://user2016.org/>

<http://user2016.org/files/abs-book.pdf>

#useR2016

@user\_stanford



# Resources

<https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016>

<http://user2016.org/>

<http://user2016.org/files/abs-book.pdf>

#useR2016

@user\_stanford



Next year @UseR\_Brussels  
July 4th - 7th, 2017

# R-Initiatives

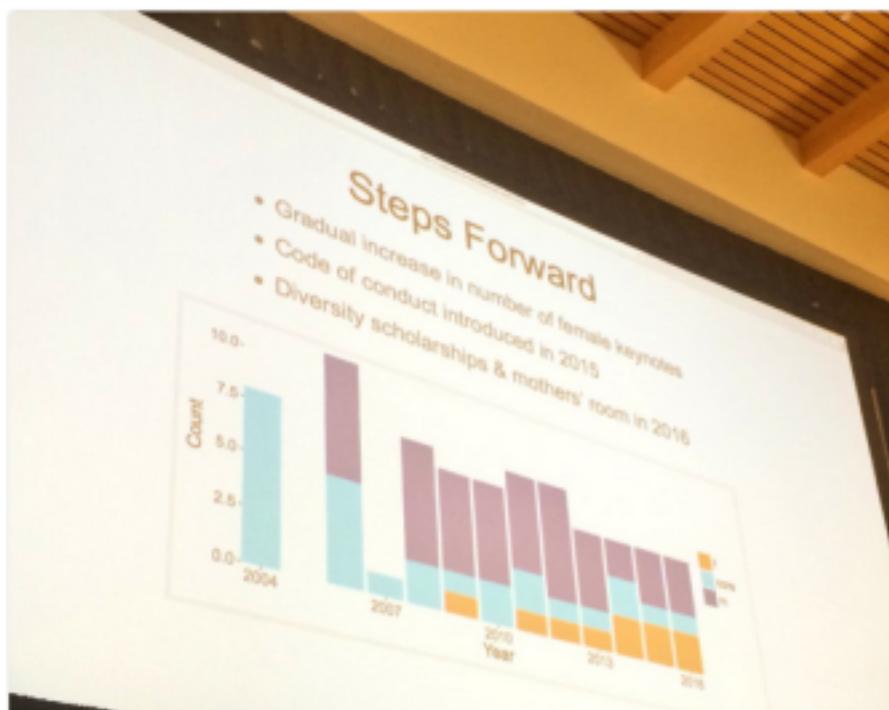
# RWomenTaskforce



Alice Data @alice\_data · Jun 27

It's not just me but @tracy73205 (my roommate from #useR2014) have noted there is definitely more women attending #useR2016 this year woop!

Progress graph by @HeathrTurnr  
@RWomenTaskforce great progress!  
#useR2016 #equality



<http://rpubs.com/hturner/useR2016>

And the stats are in #useR2016 and it's true more women this year thanks @HeathrTurnr from the @RWomenTaskforce



Myfanwy @Voovarb · Jun 27

Childcare offered at #UseR2017 = 👏 #user2016



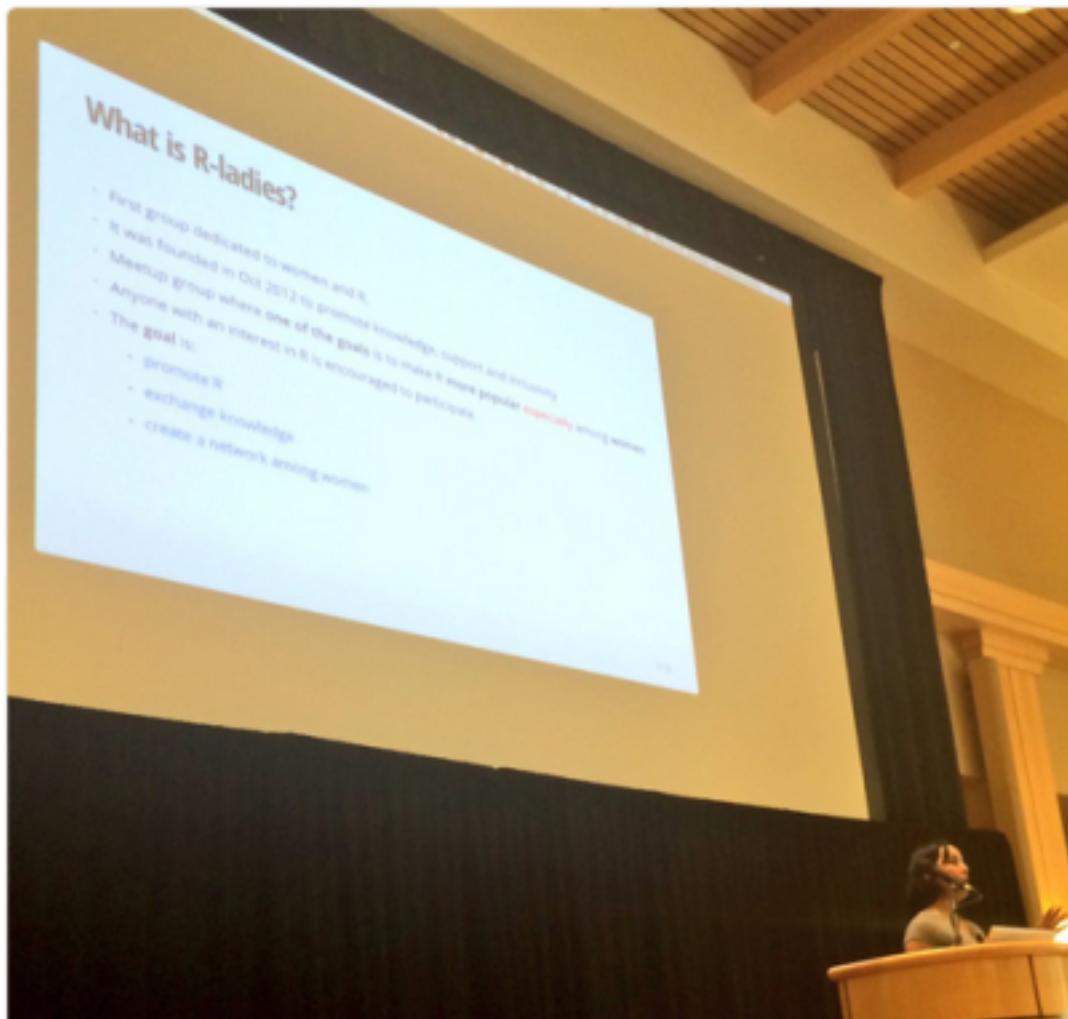
7

13

...

# San Francisco Sisters

Why was @RladiesSF was set up? They now have 700 #rladies



# R-Ladies London presentation



R-Ladies Coding Club @RLadiesLondon · Jun 27

At @useR\_Stanford? Check out @RLadiesLondon introducing ourselves to the #rstats community! [schedule.useR2016.org/event/7BVFrl...](https://schedule.useR2016.org/event/7BVFrl...)

You and Hannah Frick

27.6.16

USE R 2016 CONFERENCE

## R-Ladies London

Presentation of the Women in R Taskforce  
Alice Daish & Hannah Frick

[meetup.com/R-Ladies-Coding-Club-London/](https://meetup.com/R-Ladies-Coding-Club-London/)  
[github.com/RLadiesCodingLondon](https://github.com/RLadiesCodingLondon)  
[@RLadiesLondon](https://twitter.com/RLadiesLondon)



Matt Ritchie @mitchieau · Jun 28

@RLadiesLondon, 4 months old has 400+ members, 40+ participants per event. Tour de #Rstats a popular beginner workshop #useR2016



Scott Chamberlain @scottie · Jun 28

#useR2016 fulfill programming potential by @RLadiesLondon



SanghaChick @SanghaChick · Jun 28

@RLadiesLondon got me all inspired. #rstats #useR2016



# Great Response, and What can we do?

Lou Bajuk-Yorgan @LouBajuk · Jun 28

Kudos to all the great work being done by @RLadiesLondon #useR2016



Karl Broman @kwbroman · Jun 28

"what can we do?"

- "actively encourage women to program"
- "better job descriptions – no 'rock star'"
- "\*Do\* talk technical" #UseR2016



13

34

...

Mine CetinkayaRundel @minebocek · Jun 27

Inspired by the RLadies talks at #useR2016, anyone who wants to join forces to get RLadies RTP off the ground? #rstats



15

23

...

# Next Steps

- R-Ladies London
  - R-Ladies SF
  - Women in R Taskforce
  - Hadley Wickham
- 
- Discuss : R-Ladies the success from SF and LDN

# Next Steps

- R-Ladies London
- R-Ladies SF
- Women in R Taskforce



**Project title : “R-Ladies Alignment and Global Expansion”**

**Funding request : \$10,000**

# R-Ladies Global

## Alignment

- Mission Statement, Organisational Structure, Code of Conduct, Logo, Github, Slack
- R-Ladies global website

## Expansion

- “R-Ladies in a box”
- US: New York, Boston, Houston, Research Triangle Park
- UK: Manchester, Bristol, Scotland
- Australia: Melbourne, Sydney

# R-Ladies Global

## Alignment

- Mission Statement, Organisational Structure, Code of Conduct, Logo, Github, Slack
- R-Ladies global website

**FINGERS CROSSED!**

- "R-Ladies in a box"
- US: New York, Boston, Houston, Research Triangle Park
- UK: Manchester, Bristol, Scotland
- Australia: Melbourne, Sydney



## Women in Machine Learning and Data Science

WiMLDS is a community of women interested in machine learning and data science. We host events which include talks by prominent data scientists, lightning talks, technical workshops and networking events. Our members include engineers, technologists, statisticians, students and many other professionals who work in data science or would like to transition into this field. Our mission is to build a community of women in data science by means of connecting and educating our members with support from the Bay Area, NYC, Chicago, North Carolina and Boulder tech community. If you're interested in starting a new chapter please [reach out to us](#).

<http://wimlds.org> or [erin@wimlds.org](mailto:erin@wimlds.org)

[rladiescodingclub@gmail.com](mailto:rladiescodingclub@gmail.com)

@rladieslondon