Reproducible Research: A user's perspective on how to enable new discoveries with the OSDC

Maria Patterson, PhD
Open Science Data Cloud
Center for Data Intensive Science (CDIS)
University of Chicago

OSDC PIRE Workshop, 16 June, 2014

THE UNIVERSITY OF CHICAGO

OSDC OPEN SCIENCE DATA CLOUD

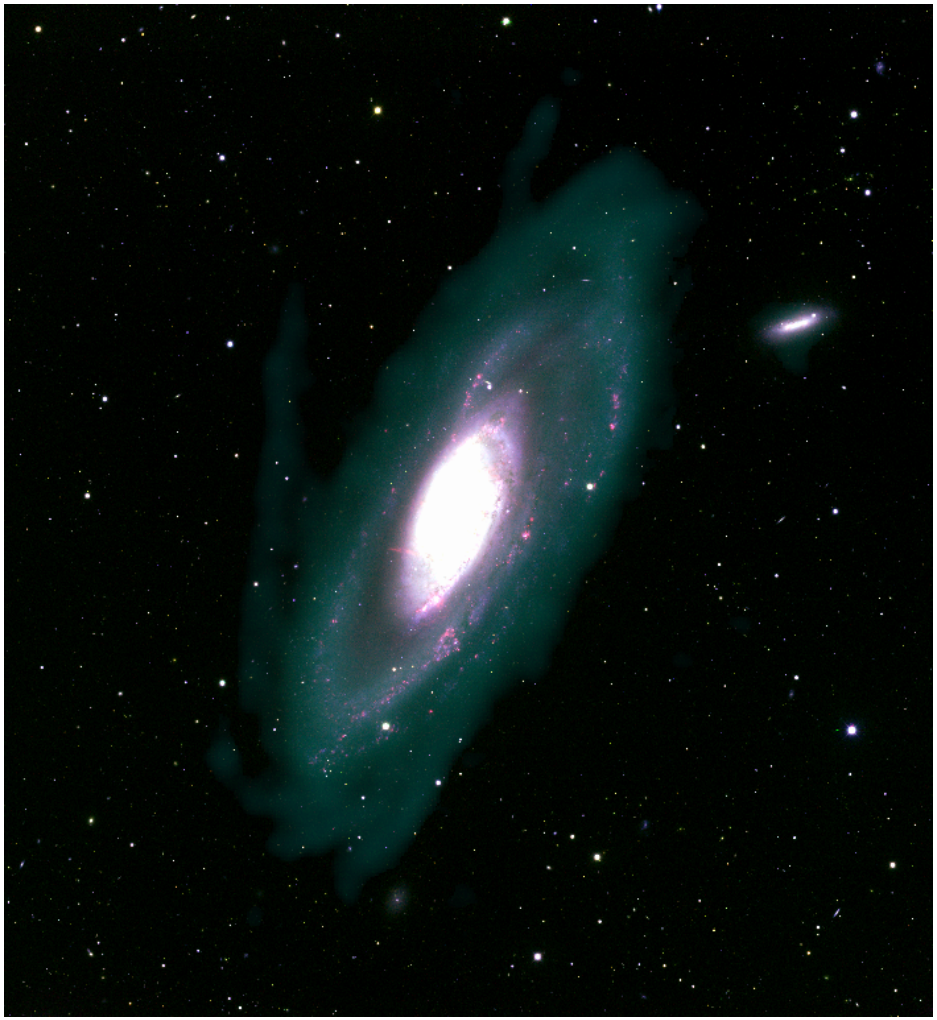
OCC OPEN CLOUD CONSORTIUM

I will talk about

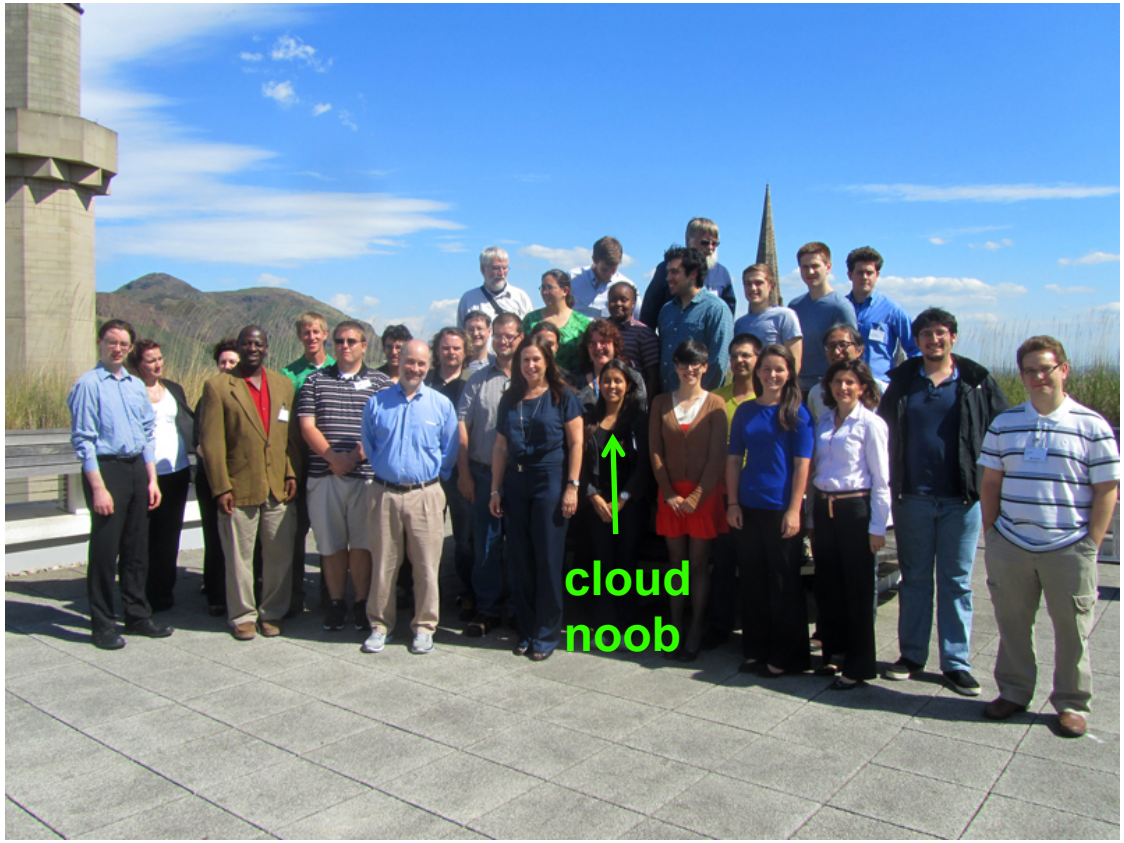
- Experience before PIRE and insight from an Astronomer
- The need for open science and reproducible research
- Tools for open and reproducible collaborative research

In grad school, I did everything wrong

- Data reduction on wide-field Mosaic tiles of optical images without using OSDC
- 282 MB x 300 images/night
  - 85 GB/night raw data
  - x4 nights run → 350 GB raw
  - x5-6 processing → ~ 2TB
- Collaborating globally with HALOGAS without using OSDC



And then I came to the 2013 PIRE workshop



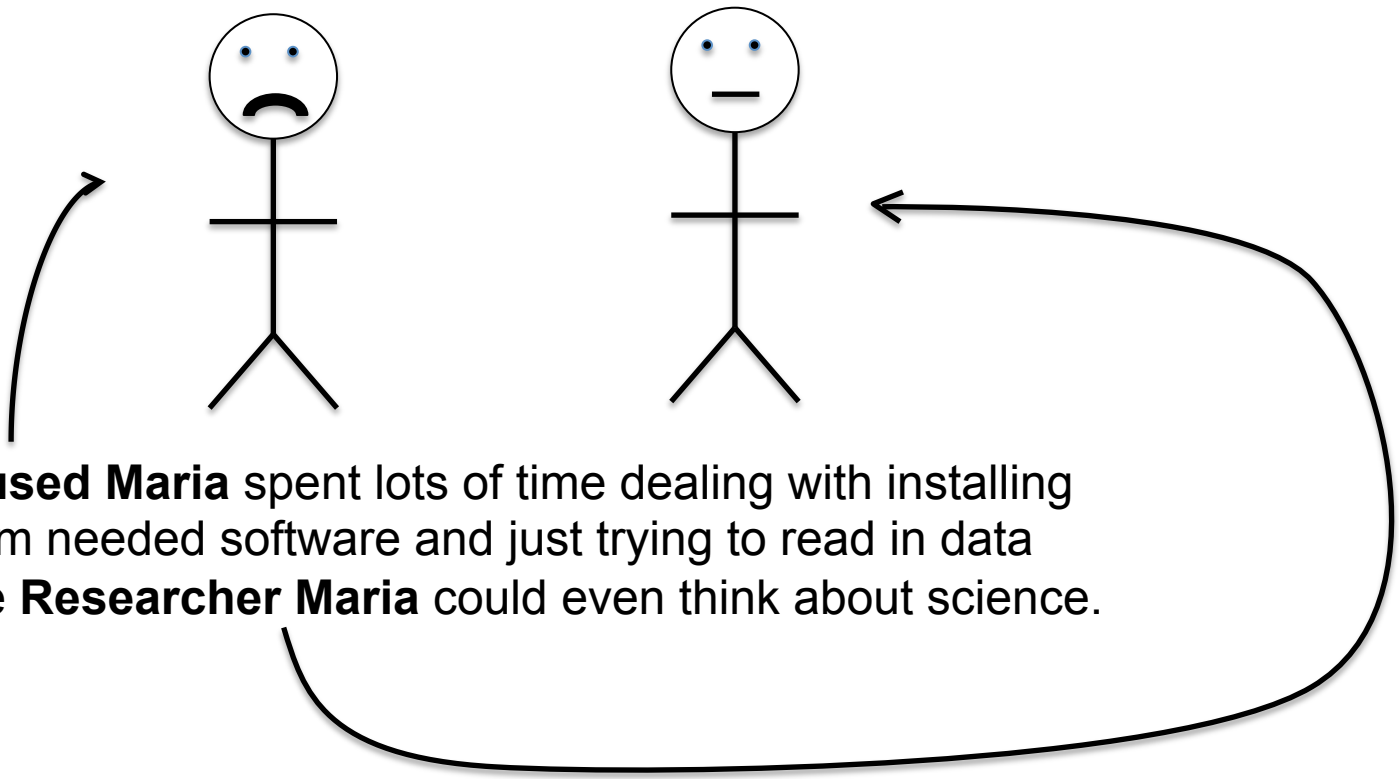
- I should have been using something like OSDC all through grad school

Astronomers have a really solid community

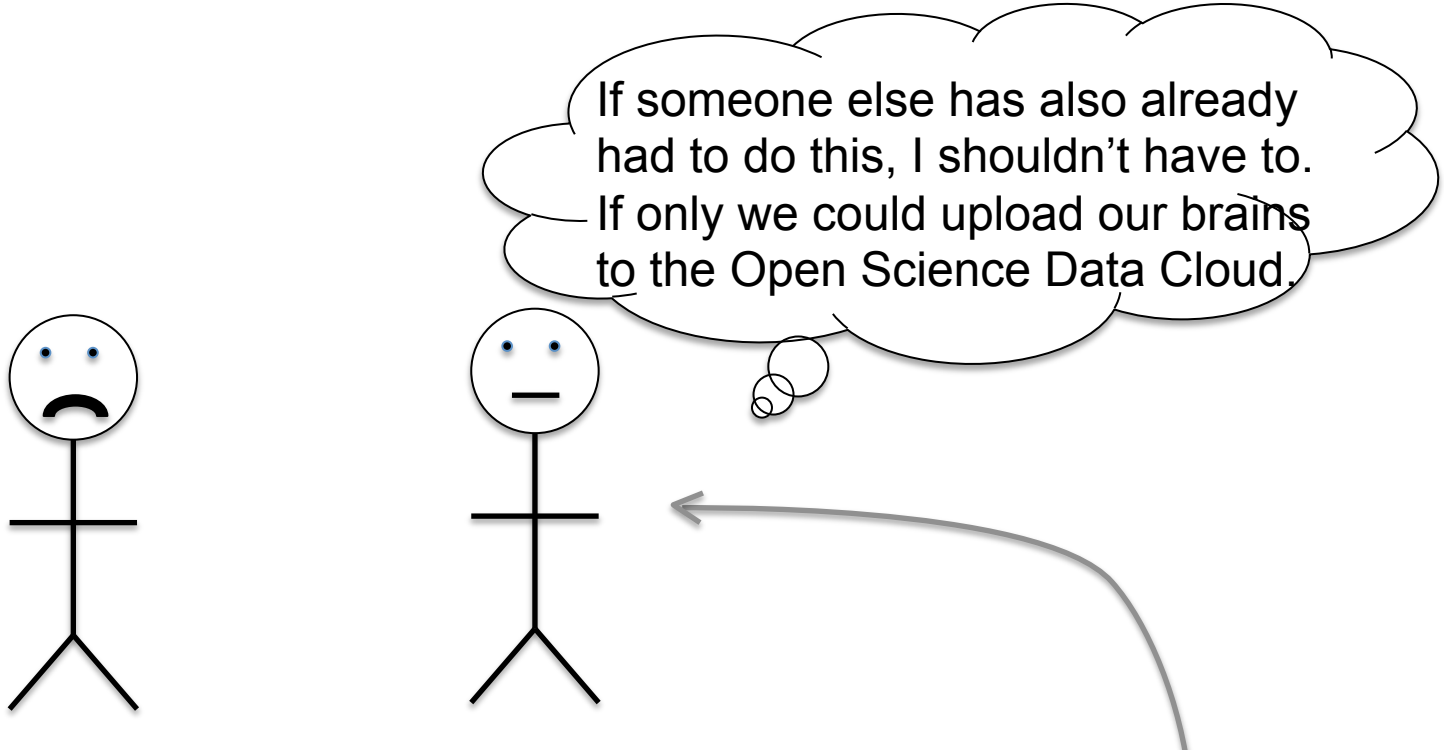
- We have pretty consistent standards.
  - .fits file format
  - IRAF software or something equivalent

- For cross-disciplinary analysis, we need a community with more cohesive data analysis approaches and collaboration to facilitate cross-disciplinary discovery.

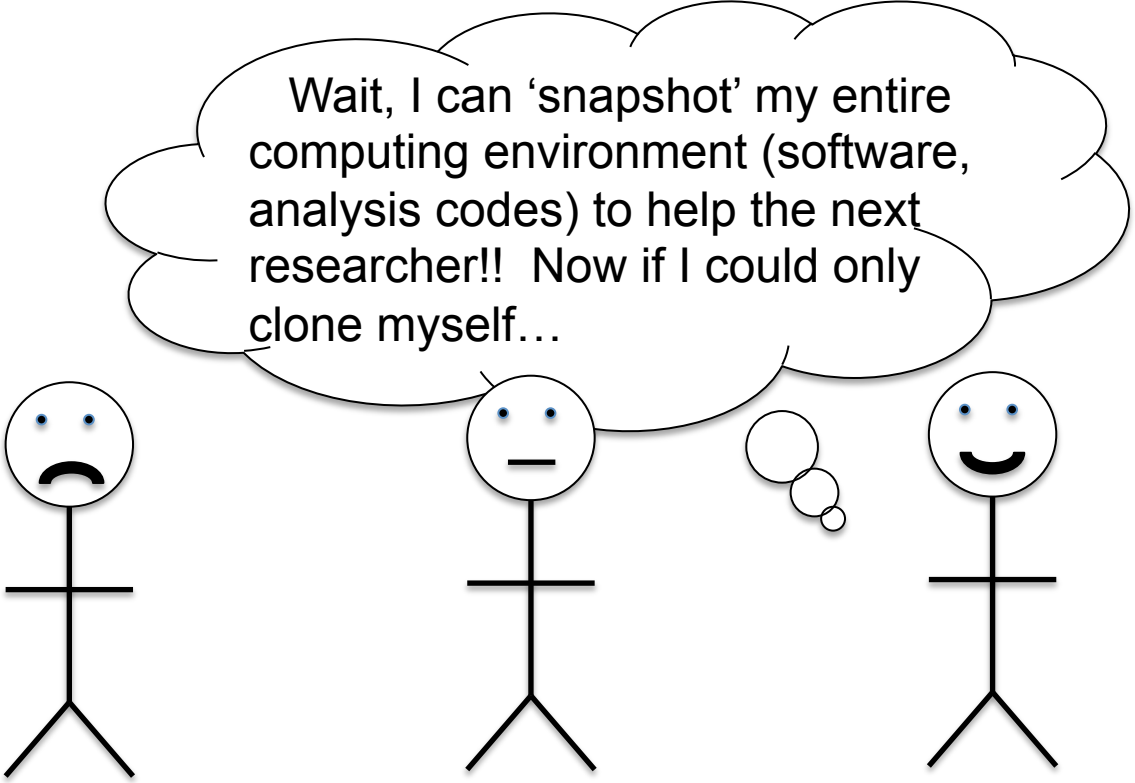
So I have been learning about How to make collaborative research easier



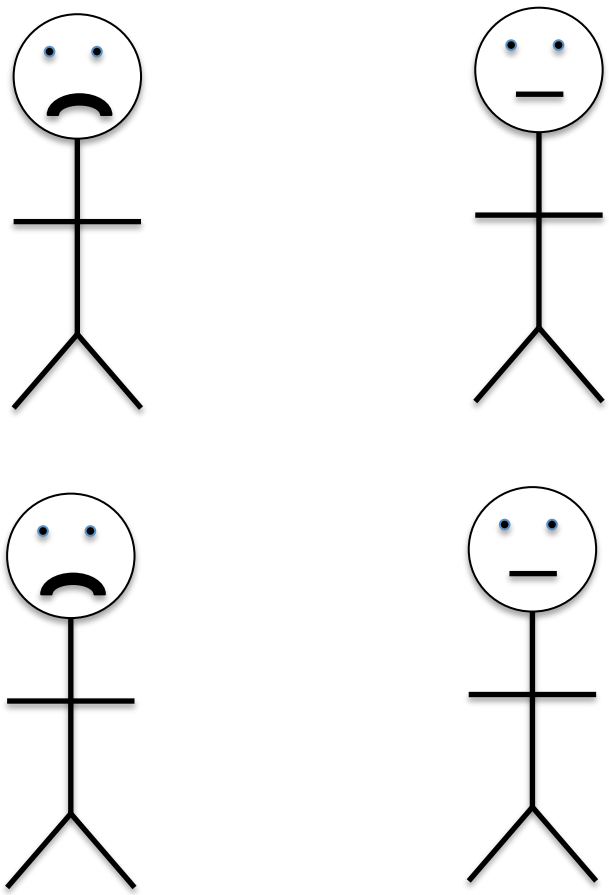
Confused Maria spent lots of time dealing with installing random needed software and just trying to read in data before **Researcher Maria** could even think about science.



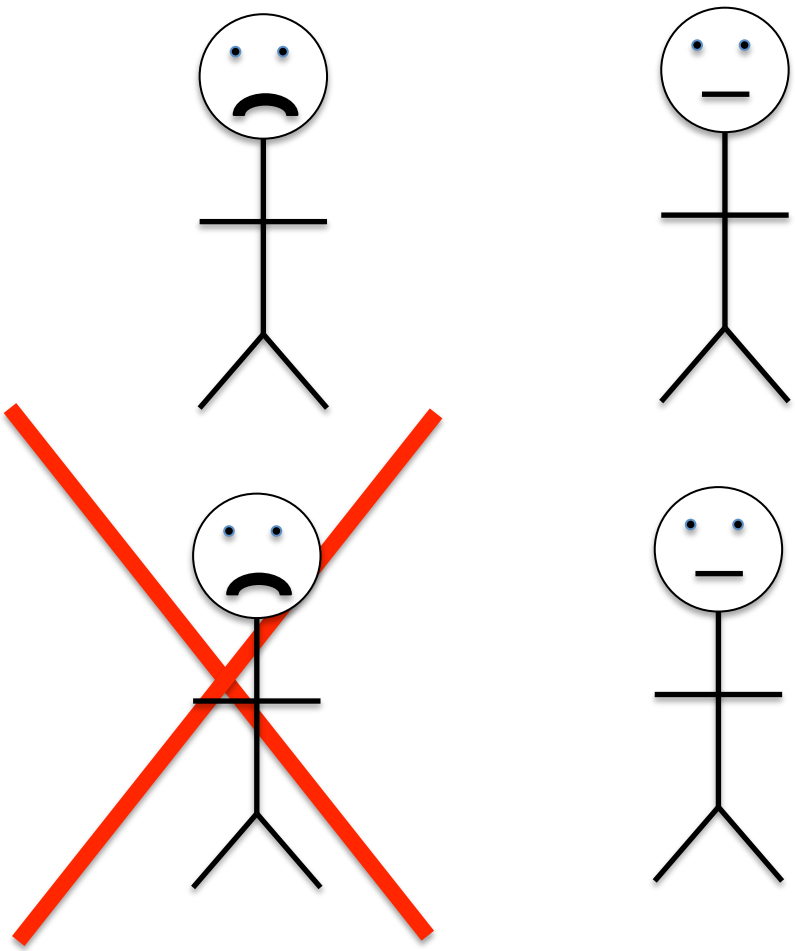
Confused Maria spent lots of time dealing with installing random needed software and just trying to read in data before **Researcher Maria** could even think about science.


Wait, I can ‘snapshot’ my entire computing environment (software, analysis codes) to help the next researcher!! Now if I could only clone myself…

*Cue **Jake the Intern**, cloned from myself*

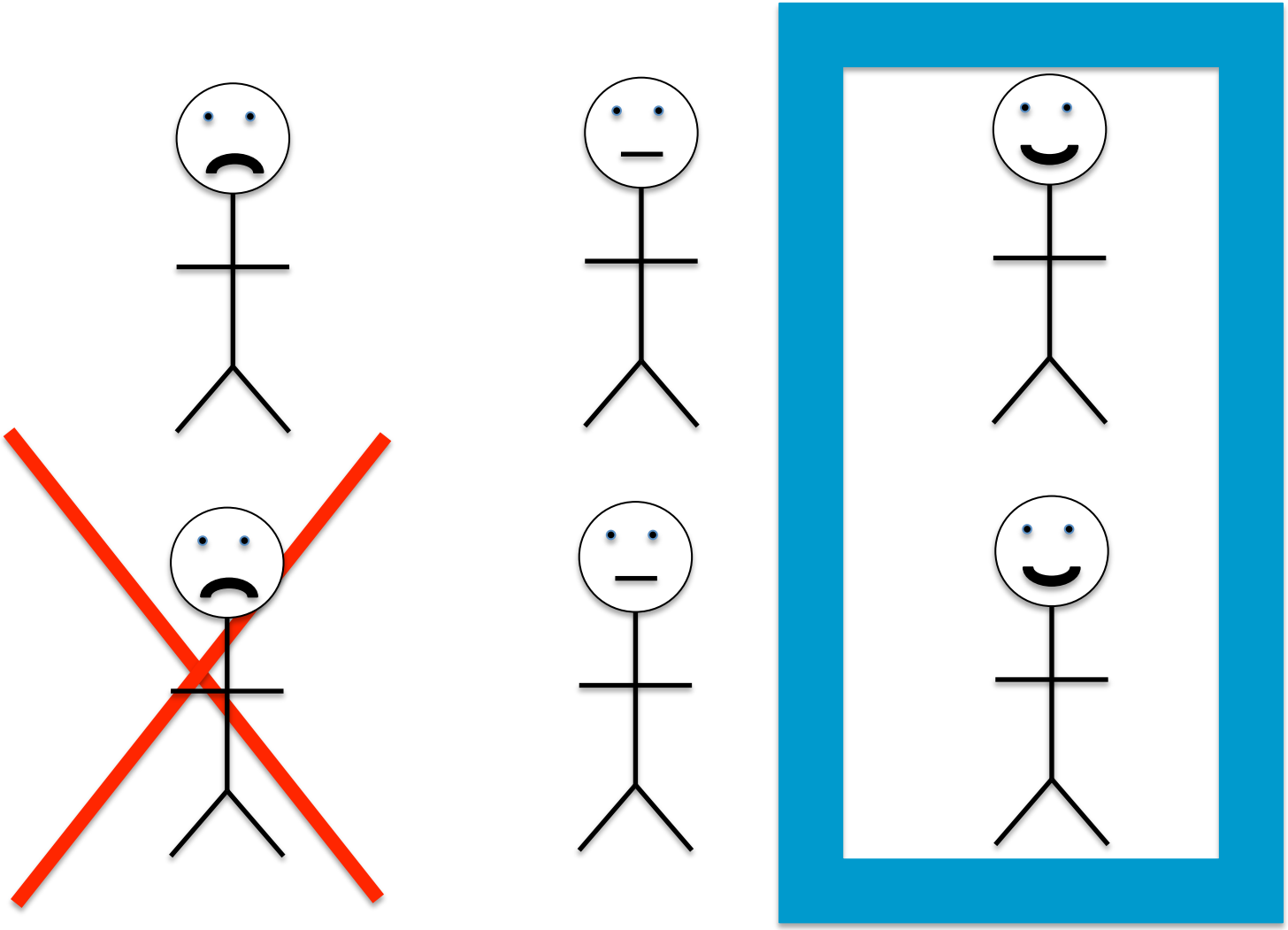


*Cue **Jake the Intern**, cloned from myself except that he never has to struggle with the stuff that confused me.*



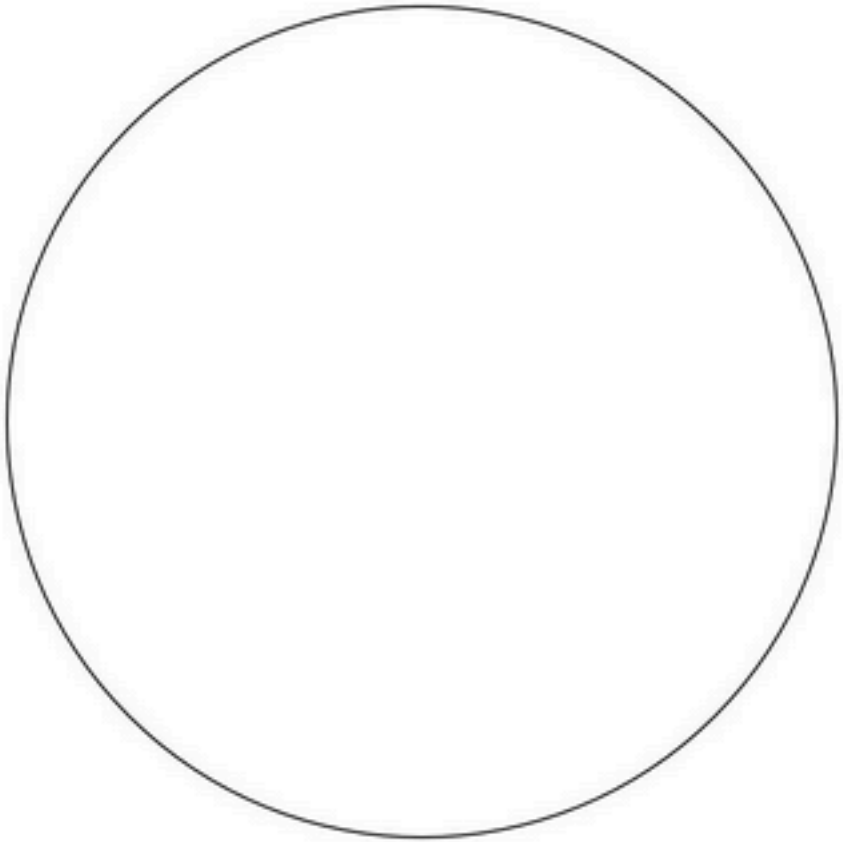
Making collaborative research easier

→ Submitted paper in < 7 weeks after Jake started



Because scientific progress is slowed redoing everyone else's work

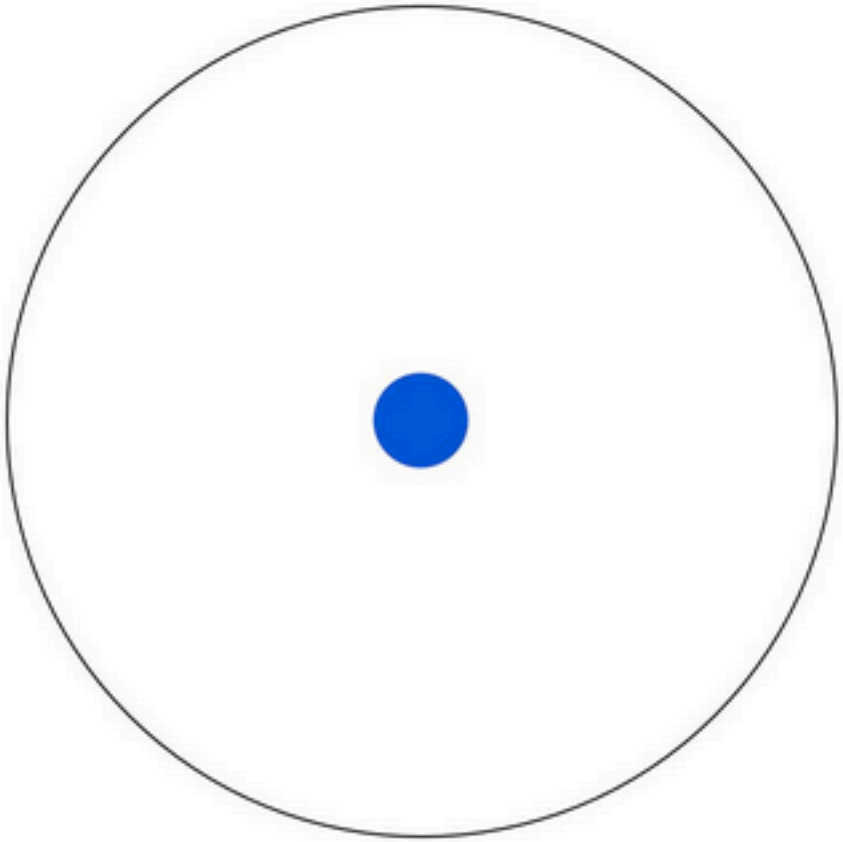
Imagine a circle that contains all of human knowledge:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

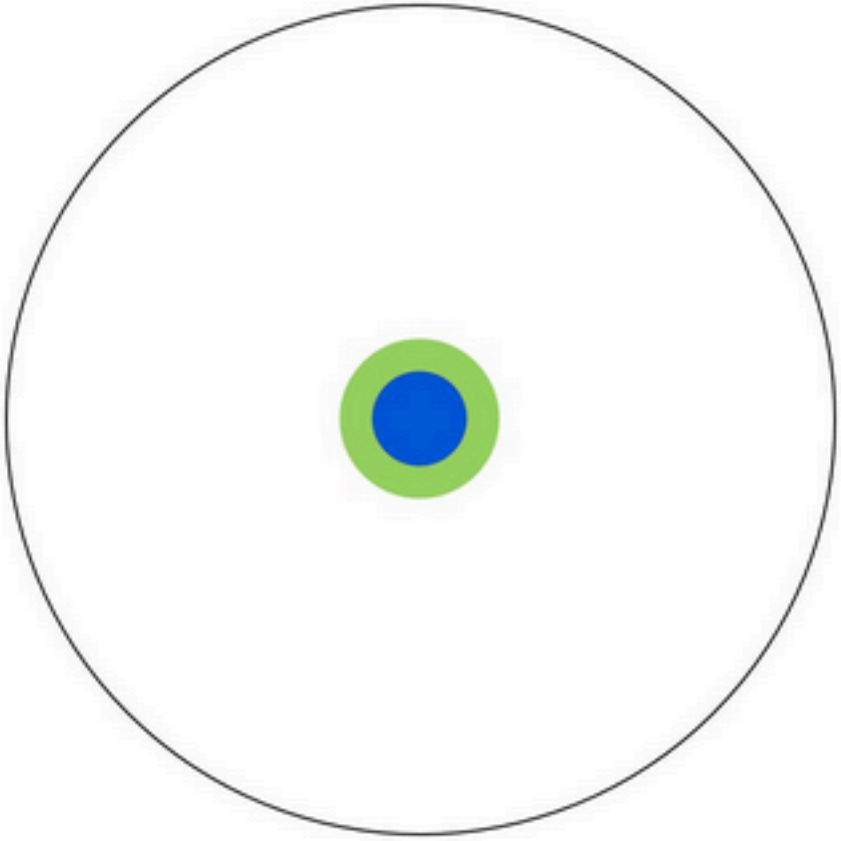
By the time you finish elementary school, you know a little:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

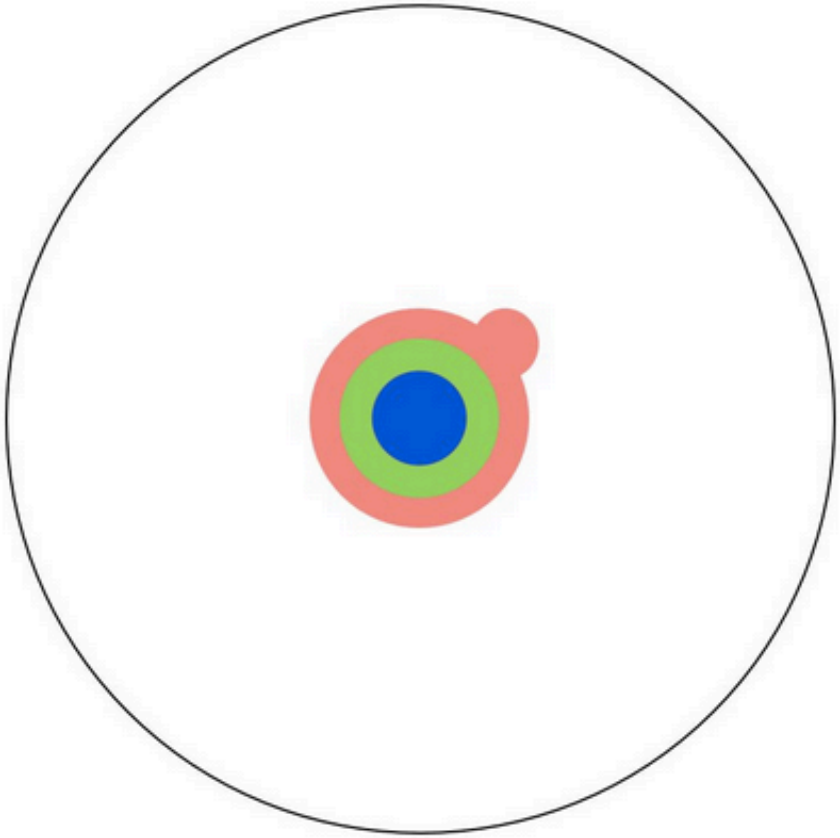
By the time you finish high school, you know a bit more:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

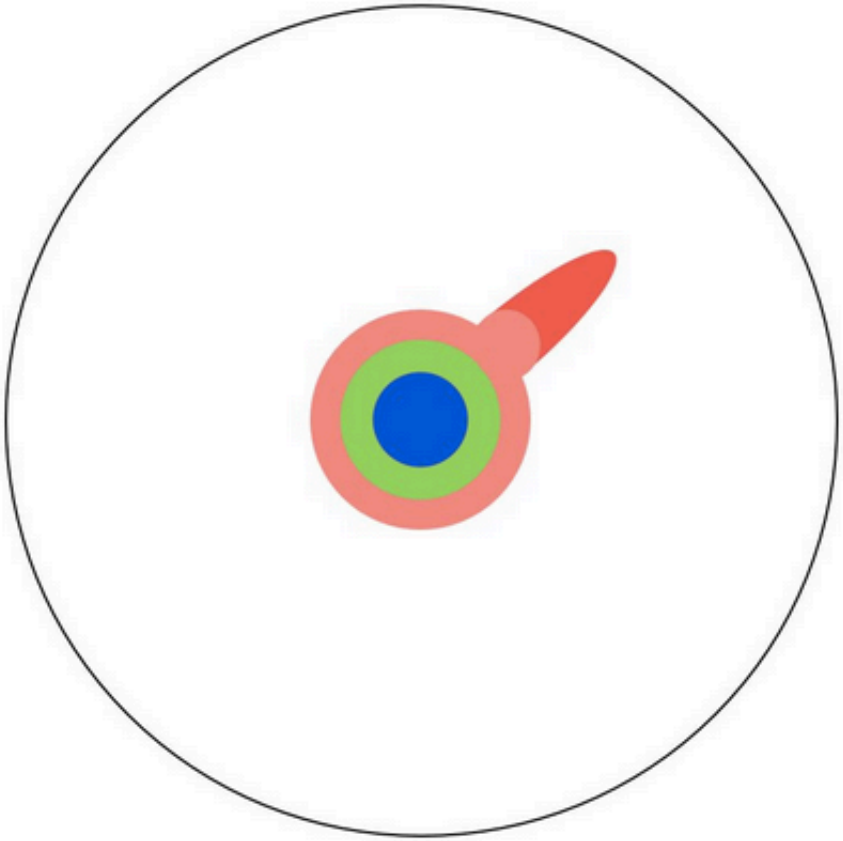
With a bachelor's degree, you gain a specialty:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

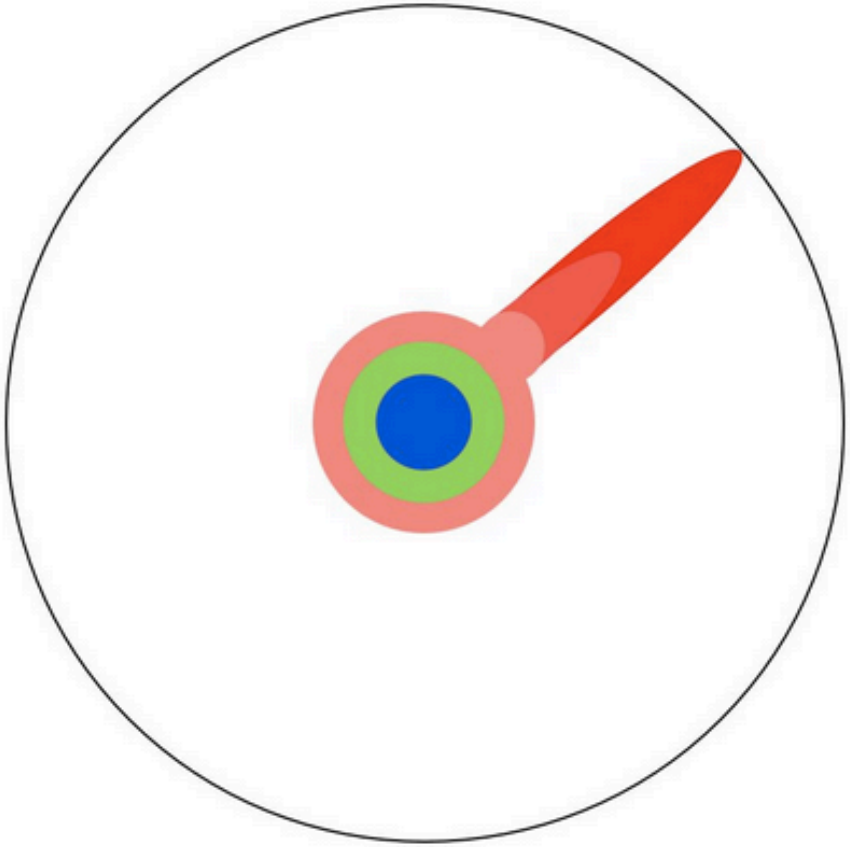
A master's degree deepens that specialty:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

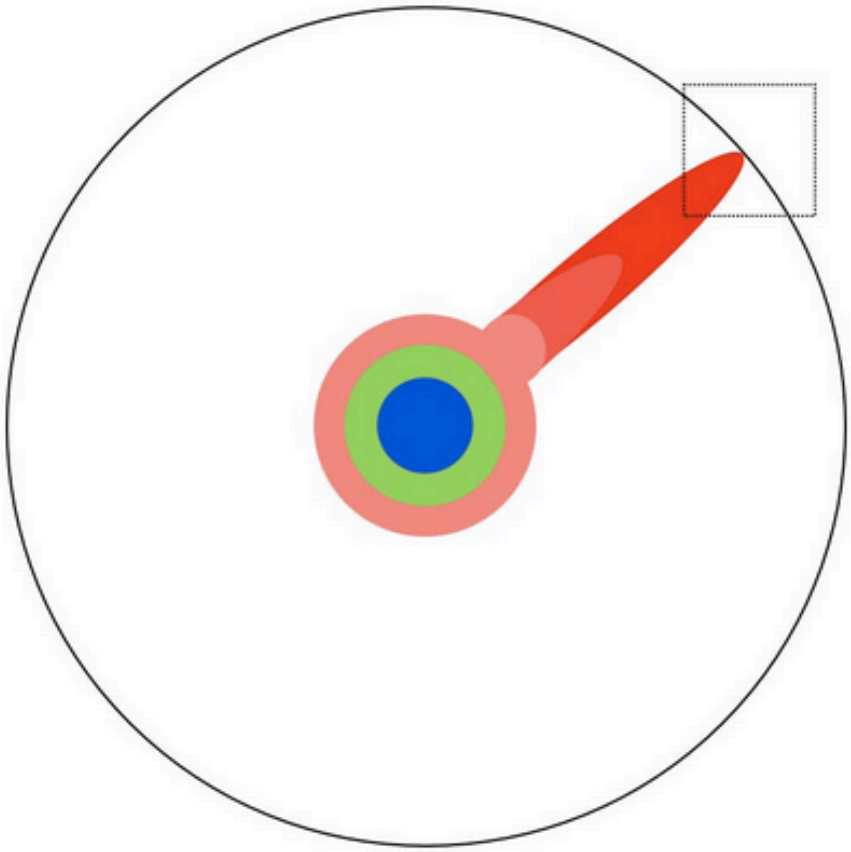
Reading research papers takes you to the edge of human knowledge:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

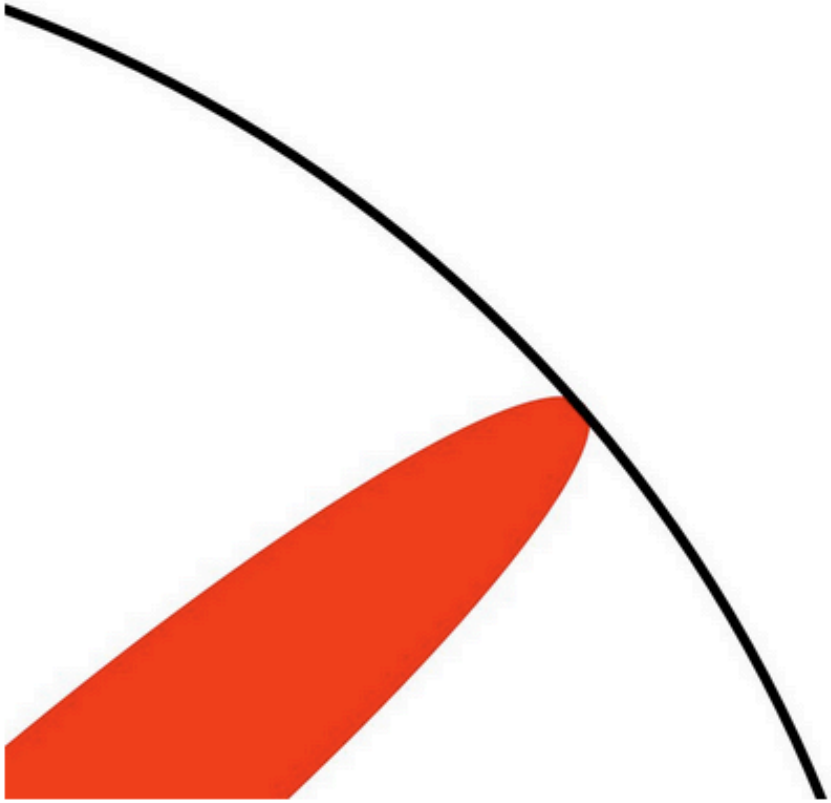
Once you're at the boundary, you focus:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

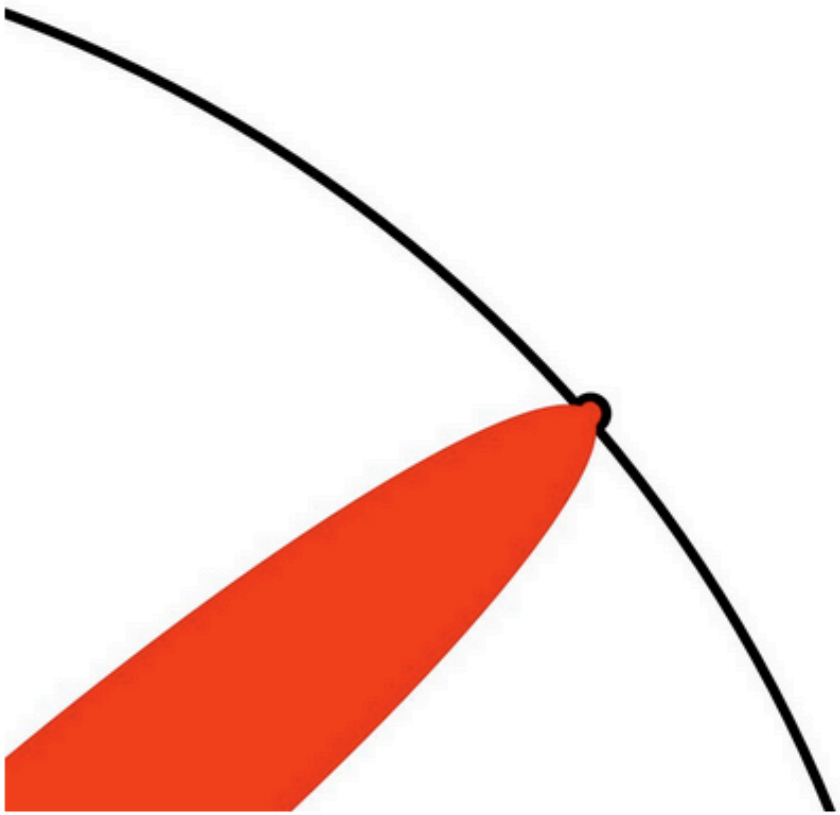
You push at the boundary for a few years:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

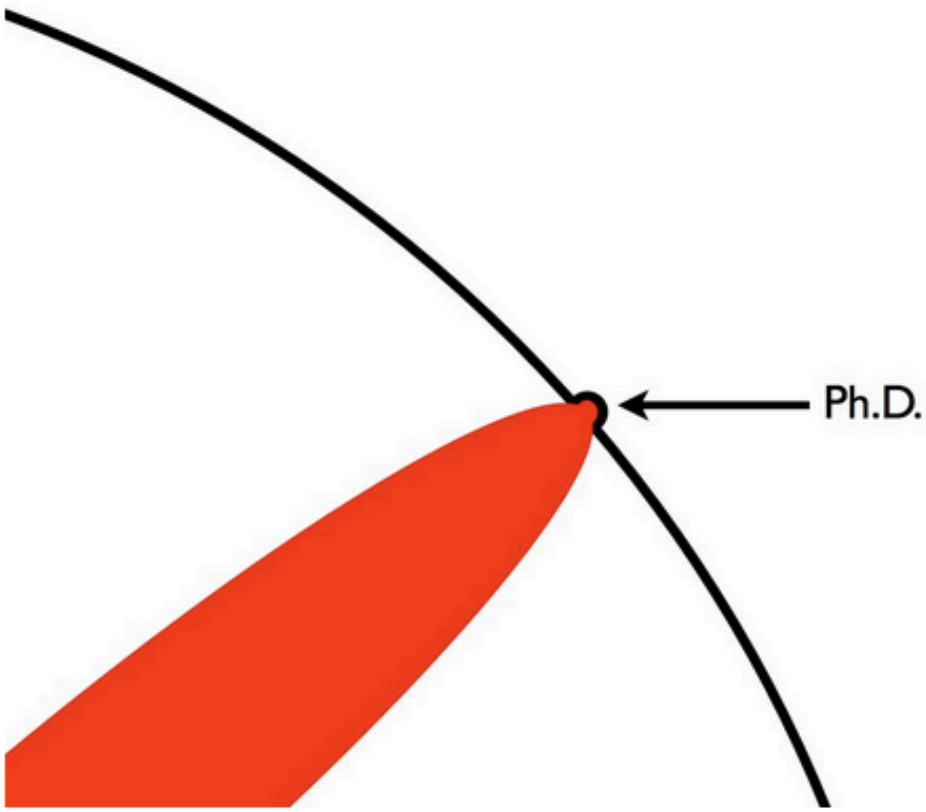
Until one day, the boundary gives way:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

And, that dent you've made is called a Ph.D.:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

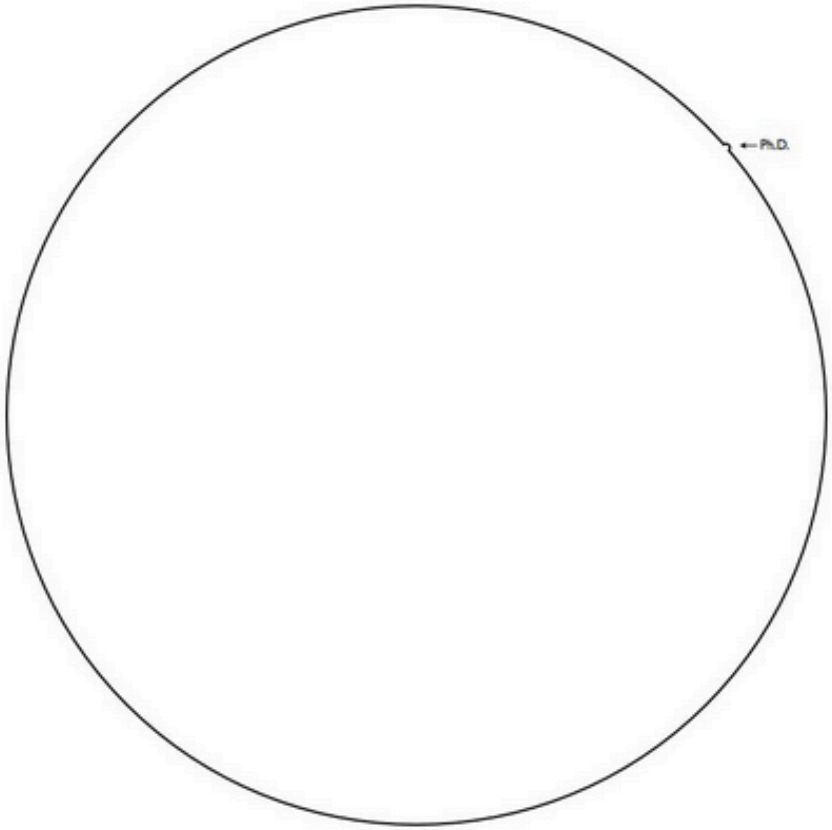
Because scientific progress is slowed redoing everyone else's work

Of course, the world looks different to you now:

Attribution : The Illustrated Guide to a PhD By Matt Might

Because scientific progress is slowed redoing everyone else's work

So, don't forget the bigger picture:



Attribution : [The Illustrated Guide to a PhD](#) By [Matt Might](#)

Because scientific progress is slowed redoing everyone else's work

Keep pushing.

Attribution : The Illustrated Guide to a PhD By Matt Might

More gravely, there is a need for reproducibility

Research: Uncovering misconduct

Virginia Gewin

Nature 485, 137-139 (2012) doi:10.1038/nj7396-137a

Published online 02 May 2012

This article was originally published in the journal *Nature*

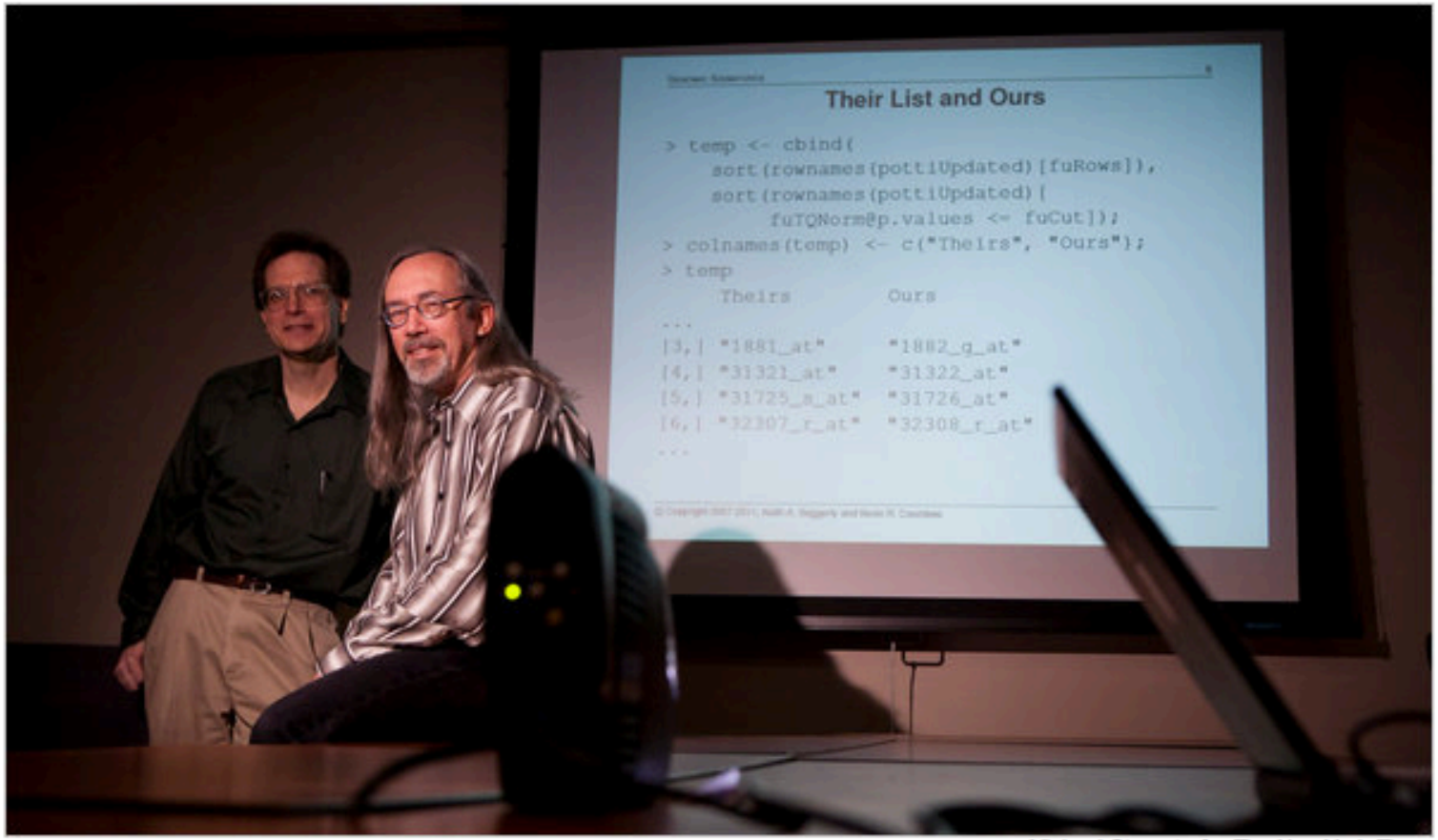
Cases of scientific wrongdoing seem to be rising. But when should researchers blow the whistle?

Biostatisticians Keith Baggerly and Kevin Coombes, like many, were intrigued by claims of personalized chemotherapy treatments by geneticist Anil Potti in 2006. Then at Duke University in Durham, North Carolina, Potti published results indicating that gene expression signatures could identify which chemotherapy drug could best treat lung or breast cancer — results that led to the setup of three clinical trials. But Baggerly and Coombes quickly found something amiss in the data. What began as concerns over apparent errors, including mislabelled samples and mismatched gene names, eventually snowballed into one of the most notorious cases of scientific misconduct in the United States in recent years.

During some 1,500 hours of work over four years, Baggerly and Coombes, both of the University of Texas MD Anderson Cancer Center in Houston, repeatedly showed that Potti's findings did not match the raw data. They analysed the data, had conversations with Potti and his supervisor, alerted the US National Cancer Institute to the likely mistakes and contacted the editors of the journal publishing Potti's work.

Repeated enquiries and complaints by Baggerly and Coombes led senior officials at the University of Texas to advise them to drop what was starting to look like a vendetta. "We were focused on the fact that the data used to justify clinical trials were wrong; we thought that should be enough," says Baggerly. "How the data got in this shape was not our immediate concern."

How Bright Promise in Cancer Testing Fell Apart



Michael Stravato for The New York Times

Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors.

By GINA KOLATA
Published: July 7, 2011

The Importance of Reproducibility in High-throughput Biology by Dr. Keith Baggerly

A solution is reproducible research

AIP | Computing in Science & Engineering

Reproducible Research

Sergey Fomel[1] and Jon F. Claerbout[2]

+ VIEW AFFILIATIONS

Comput. Sci. Eng. **11**, 5 (2009); http://dx.doi.org/10.1109/MCSE.2009.14

< PREVIOUS ARTICLE | TABLE OF CONTENTS | NEXT ARTICLE >

Abstract | Full Text | References (0) | Cited By | Data & Media | Metrics | Related

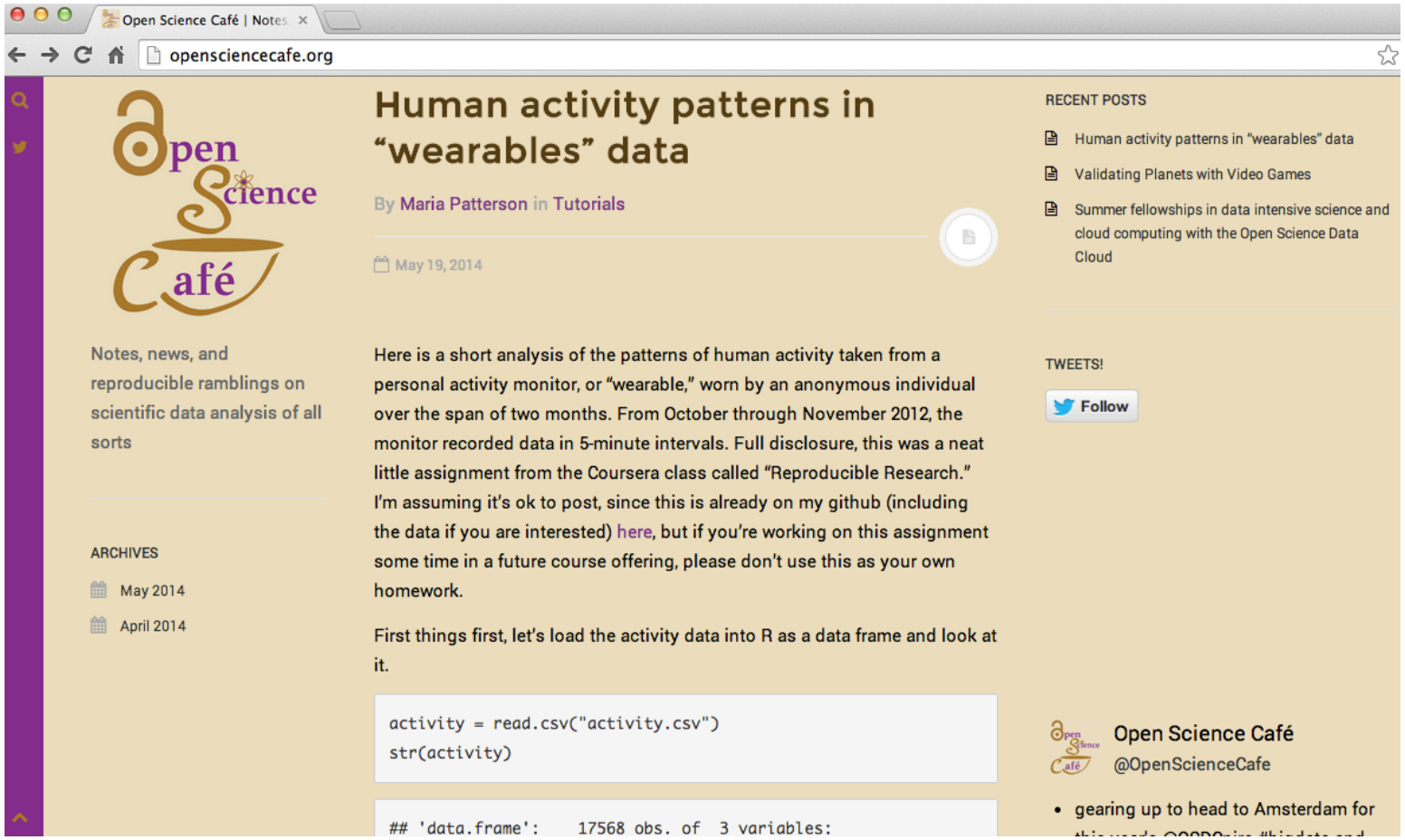
Reproducibility is a core principle of science. For computational experiments to become reproducible, one needs to develop a system for linking scientific publications with computational recipes. Articles in this special issue argue in favor of computational reproducibility and describe several practical approaches to reproducible research.

What is reproducible research?

- Making data analysis transparent by allowing others to completely reproduce all results from the initial raw data
  - Same data, same computing environment, same analysis

- Can do this with Open Science Data Cloud

This slide is shameless self promotion



Dr. Jeff Coughlin (Seti Institute, Kepler), Dr. Ryan Hamilton (NASA Ames, SOFIA), what a great blog this should be

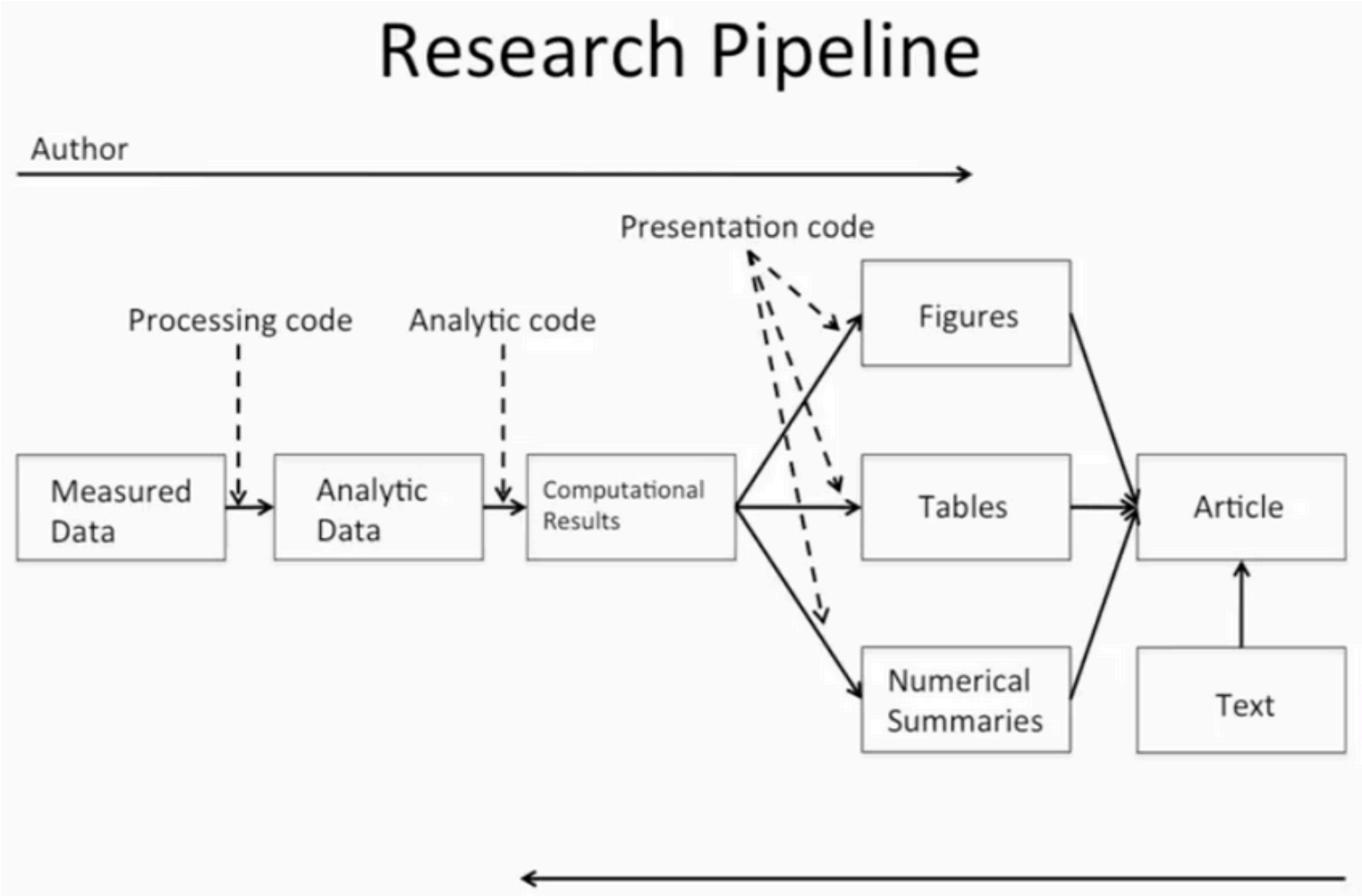


Figure source: Roger Peng, JHU
Reproducible Research class, Coursera

Literate statistical programming is a great tool

- Programming that is both
  - Human readable
  - Machine readable

- Embed your code in the analysis documentation

Some tools for reproducible research: Sweave

```
\documentclass[a4paper]{article}

\title{Sweave Example 1}
\author{Friedrich Leisch}

\begin{document}

\maketitle

In this example we embed parts of the examples from the
\texttt{kruskal.test} help page into a \LaTeX{} document:

<<>>=
data(airquality)
library(ctest)
kruskal.test(Ozone ~ Month, data = airquality)
@
which shows that the location parameter of the Ozone
distribution varies significantly from month to month. Finally we
include a boxplot of the data:

\begin{center}
<<fig=TRUE,echo=FALSE>>=
boxplot(Ozone ~ Month, data = airquality)
@
\end{center}

\end{document}
```

Figure 1: A minimal Sweave file: `example-1.Snw`.

Sweave Example 1

Friedrich Leisch

May 15, 2008

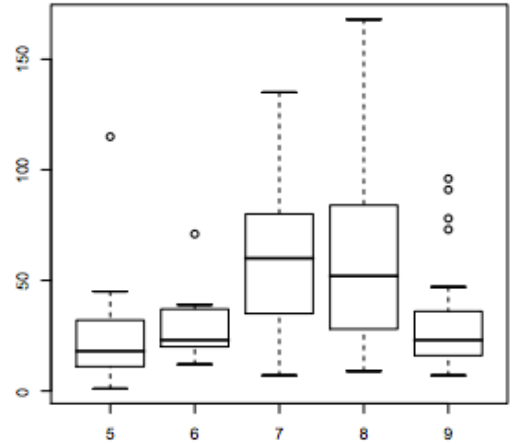
In this example we embed parts of the examples from the `kruskal.test` help page into a LaTeX document:

```
> data(airquality)
> library(ctest)
> kruskal.test(Ozone ~ Month, data = airquality)

        Kruskal-Wallis rank sum test

data:  Ozone by Month
Kruskal-Wallis chi-squared = 29.2666, df = 4, p-value = 6.901e-06
```

which shows that the location parameter of the Ozone distribution varies significantly from month to month. Finally we include a boxplot of the data:



Some tools for reproducible research: knitr

- R markdown ‘knitr’ demonstration (no LaTeX knowledge required)

Checklist for reproducible research

- Start with a good science question.
- Do NOT do anything by hand (edit spreadsheets, remove outliers, point and click to download data) but instead teach the computer to do it.
- Use version control.
- Set seeds for random number generation.
- Keep track of your computing environment (OS, architecture, software dependencies and versions).
- Don’t save any output (tables, figures), but instead keep the code + data to regenerate them.
- Think about the entire pipeline, from accessing raw data entirely through to displaying results.

Additional online tools for collaboration

- See writelatex

In summary

- Please take advantage of this workshop to cross-pollinate the best ideas from different disciplines.

- Use reproducible research methods to
  - 1) Make your life easier
  - 2) Push the boundary of knowledge further
  - 3) Ensure valid and reliable scientific results

- Don’t forget to tweet what you have learned and mention @OSDCpire