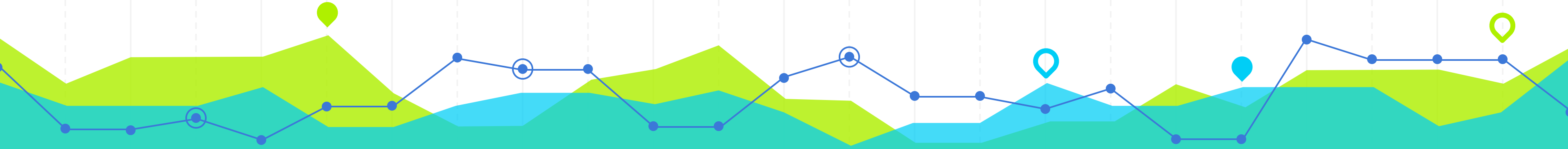


Data Science with R

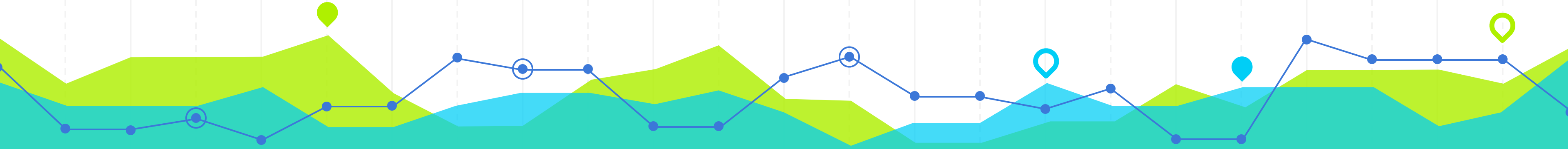
OSU Math to Industry R Workshop
September 20, 2018

HELLO!

My name is Collin McCabe

I’m a Data Scientist at Radiology Partners.

But before data science, I was an academic research scientist, and I still draw on this background daily.



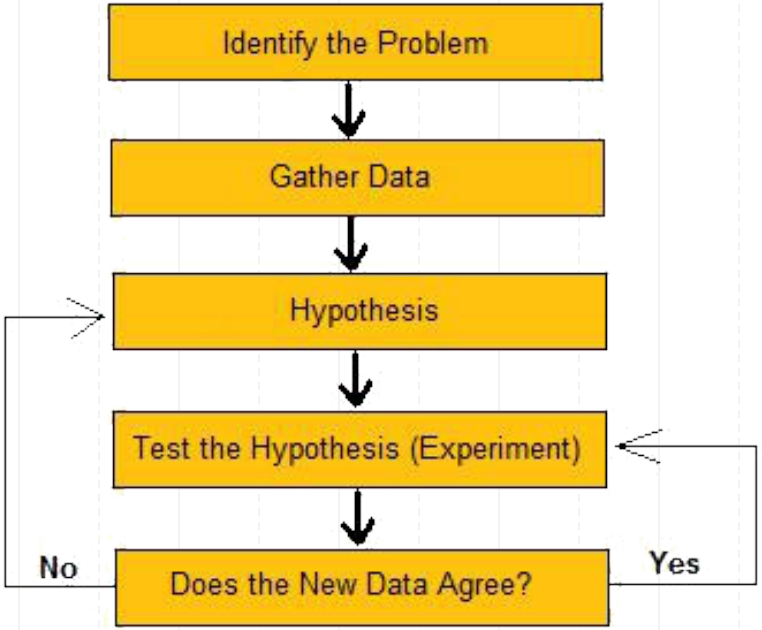
What Is Data Science?

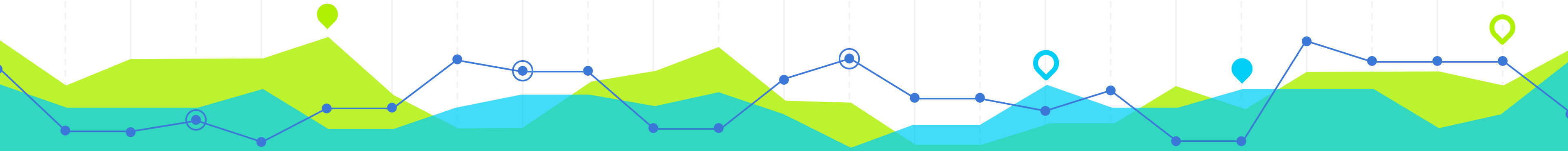
… and why is it so hot right now?

1

Data science is SCIENCE, first and foremost!

- Identify significant problems to solve & questions to ask
- Collect the data to answer your questions
- Explore data and discover patterns
- Test models to explain the data
- Communicate findings to others to incite action
- Iterate, iterate, iterate





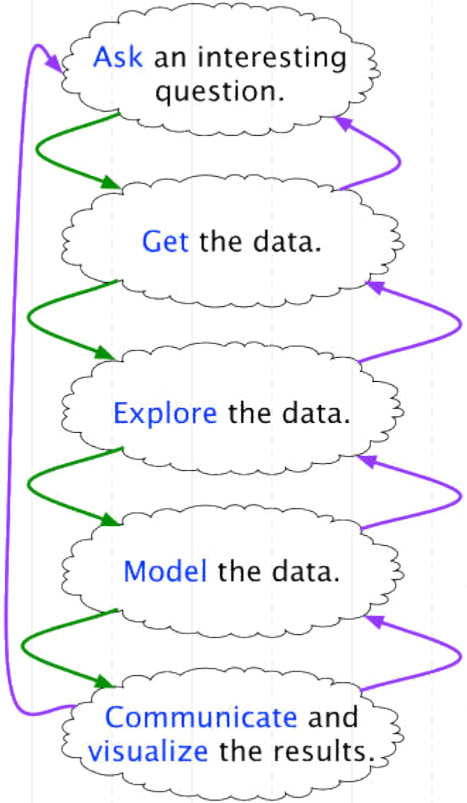
What Do Data Scientists Do?

… and how do they do it?

2

Data scientists apply the scientific method to solve open-ended, data-heavy problems

- Identify questions
  - Experience, subject matter experts (SMEs), etc
- Collect data
  - Sensors, surveys, databases (SQL, Hadoop), etc
- Clean, analyze, and model data
  - R, Python, Julia, SAS, etc
- Visualize data
  - ggplot (R), matplotlib (Python), d3.js, Tableau, etc
- Communicate findings
  - Public speaking, writing reports, blogging, etc

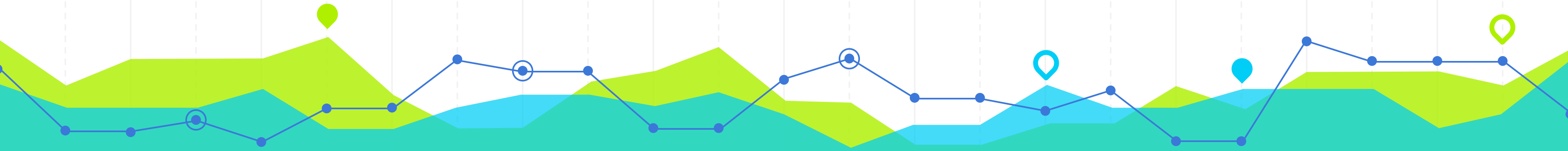

Ask an interesting question.

Get the data.

Explore the data.

Model the data.

Communicate and visualize the results.



What Is R?

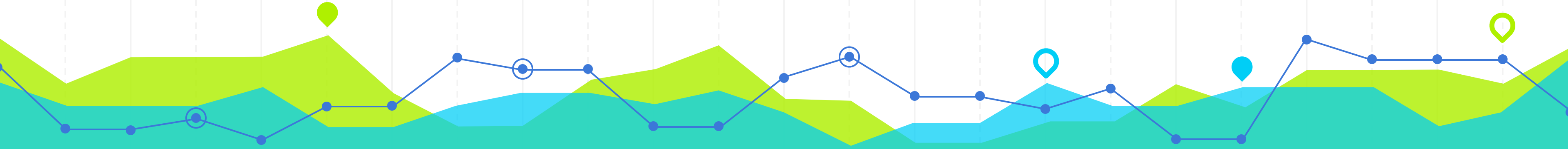
Answer: awesome.

3

> *R is a language and **environment** for **statistical computing** and **graphics**.*
>
> *- The R Foundation*

R is …

- Open source & free software
  - You never have to pay, and it respects your freedom
- Modular and package driven
  - Only comes with the functionality you need, you add to it
- Active user base
  - New packages released often, lots of help
- Command line driven
  - Learning curve, but easy to save/reproduce workflow



What Is a Statistical Computing Environment?

… and why do I need one?

4

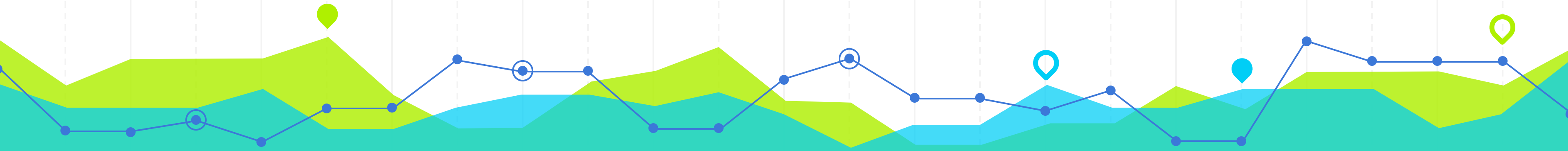
Statistical Computing Environments

Statistical Computing

- Using computer science together with mathematics and statistics
- Pushing the boundaries of knowledge by leveraging increasing computing power
- Creating reusable code and statistical methods that can easily be replicated applied by others
- Developing or utilizing algorithms and functions to automate analysis

Environment

- A fully planned and coherent system
- All parts made and intended to work well with other parts
- Addition of new packages builds upon existing capabilities to avoid conflicts and/or reinventing the wheel



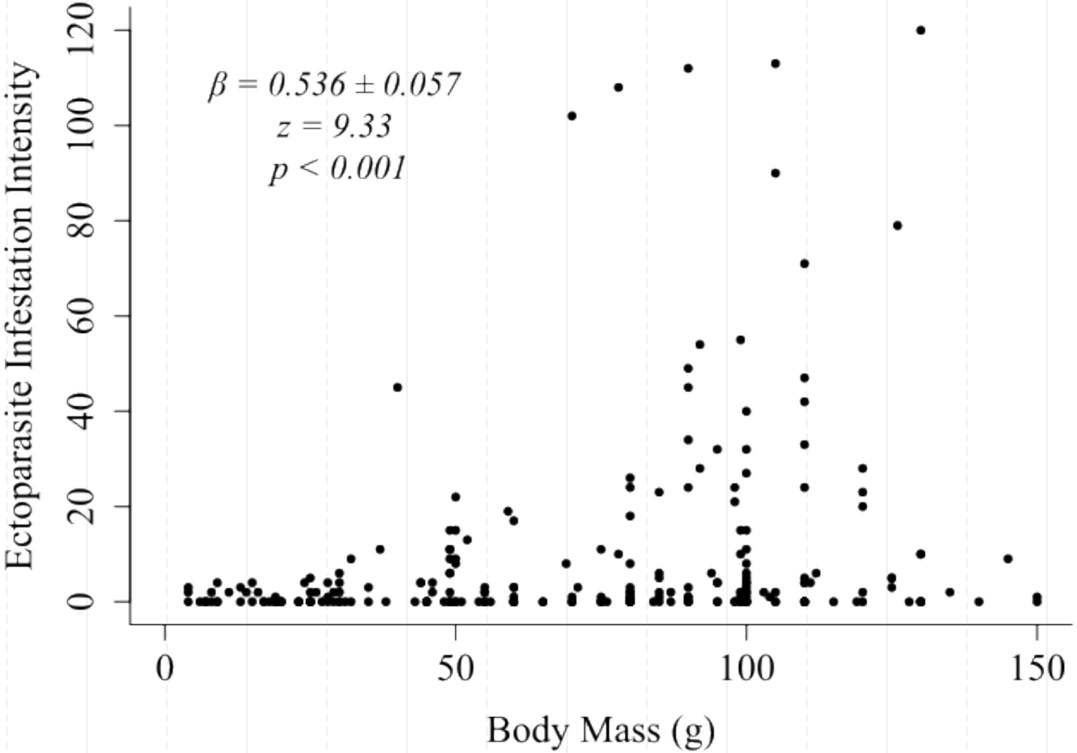
Why Program Graphics?

… aren’t my Excel graphs good enough? No.

5

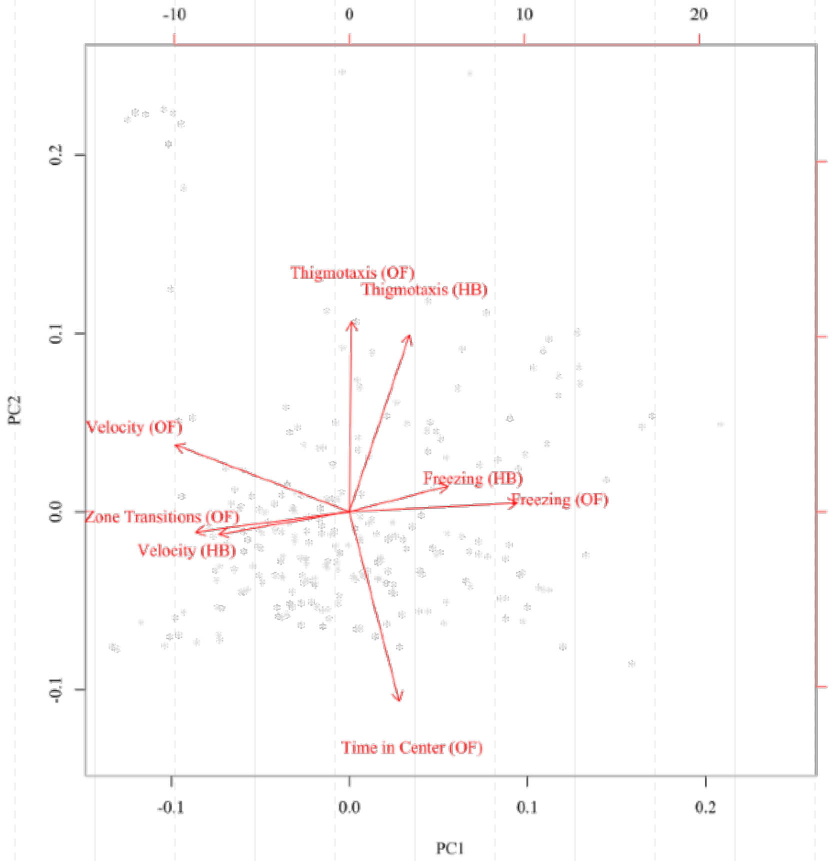
Graphics in R

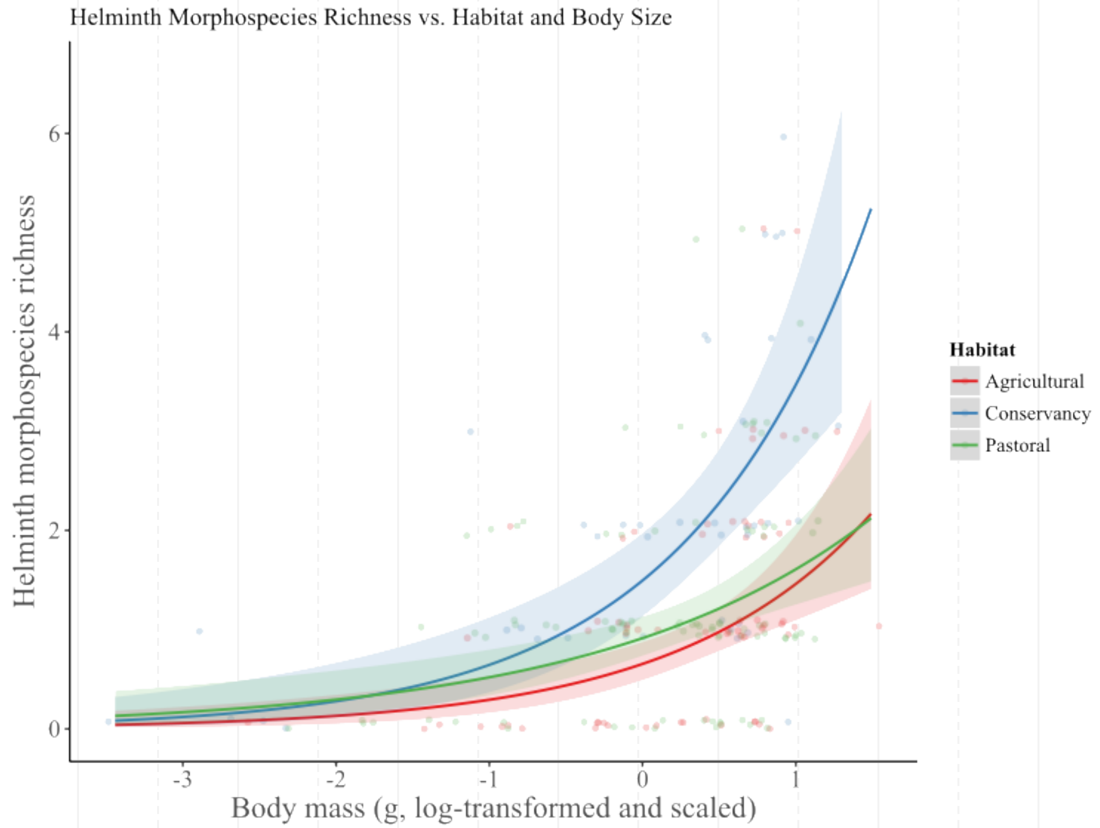
- What do you use for graphics currently?
  - Excel?
  - WYSIWYG statistical software (JMP, SPSS, Stata, etc)?
  - Adobe Illustrator (or open source alternatives)?
- Most other options are inflexible, often with ugly or unnecessary defaults that can't easily be changed
- R provides a graphics sandbox (similar to Illustrator) which can also be replicated or repurposed by sharing code
  - For example …



Graphics in R

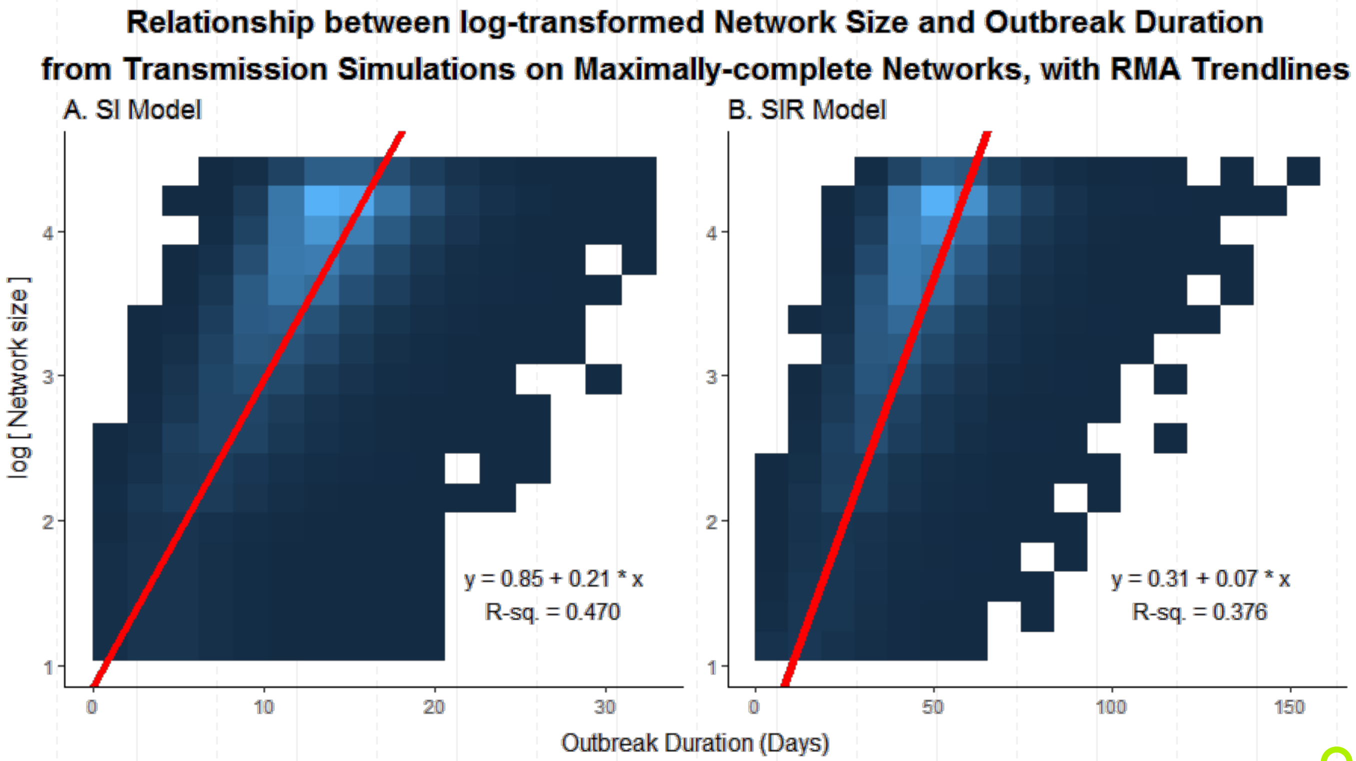
Graphics in R



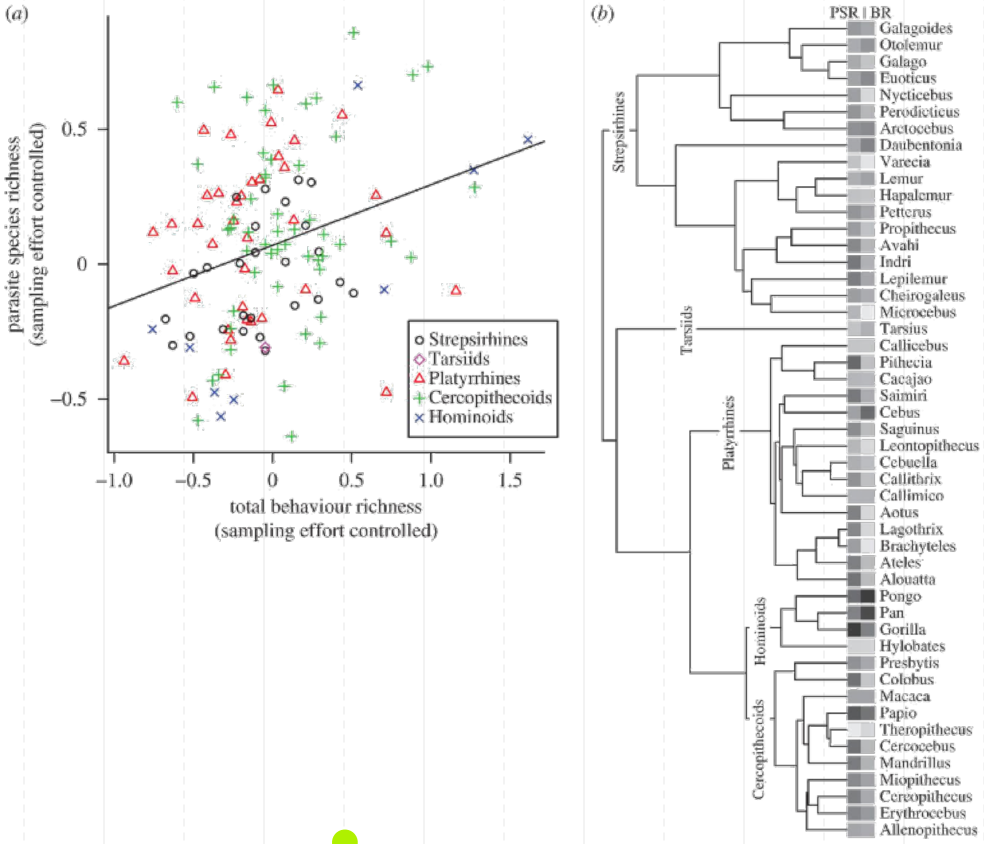


Graphics in R

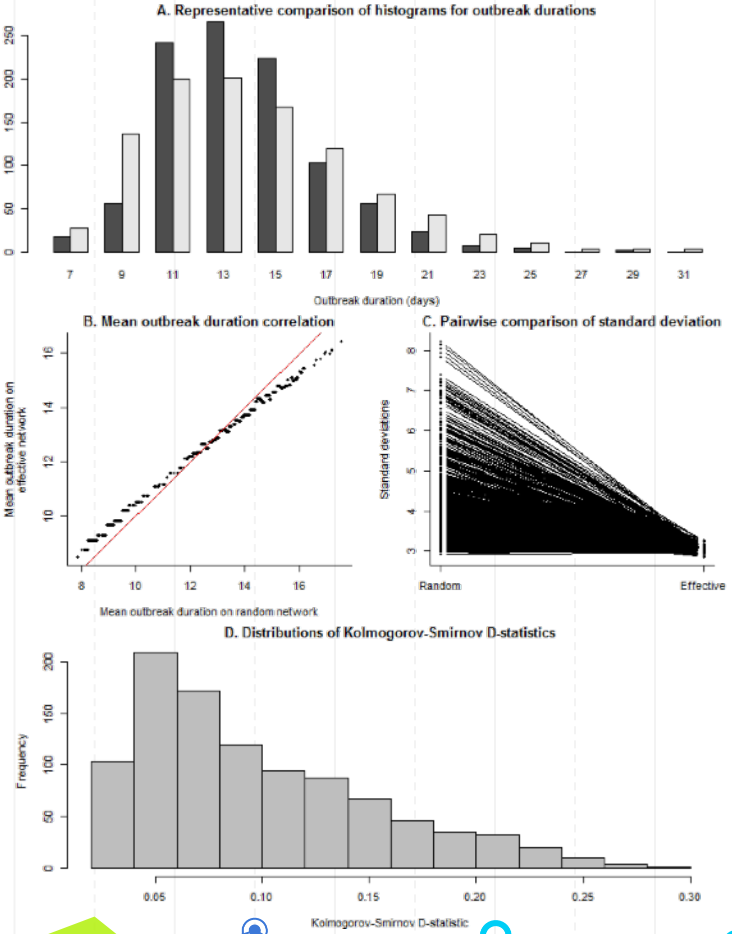
Graphics in R



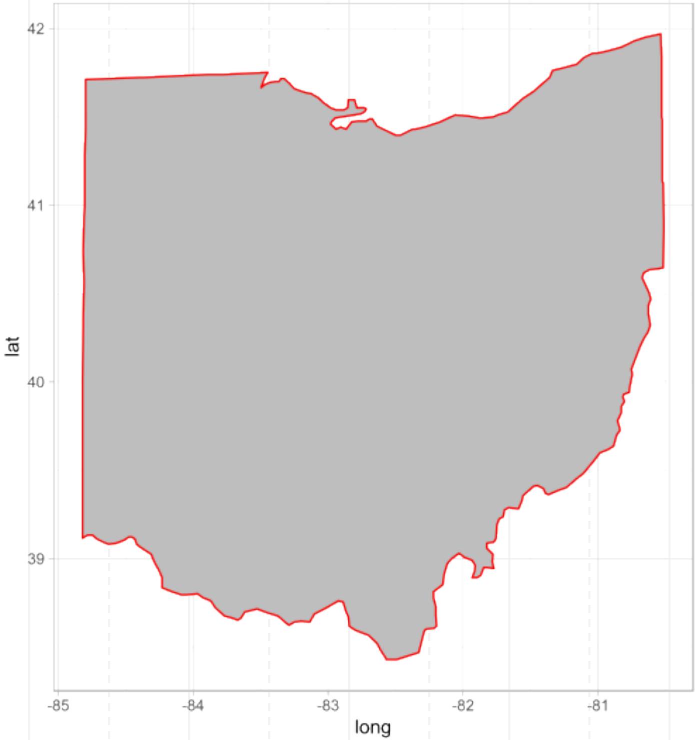
Graphics in R

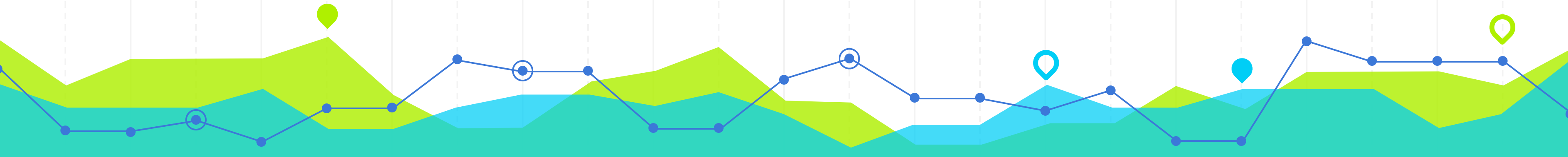


Graphics in R



Graphics in R





What About Other Languages?

… isn’t Python taking over data science?

6

How does R compare to other data science software?

SAS

- Enterprise = not free
- Large corporations use for the support and for legacy code
- Most are moving away from SAS

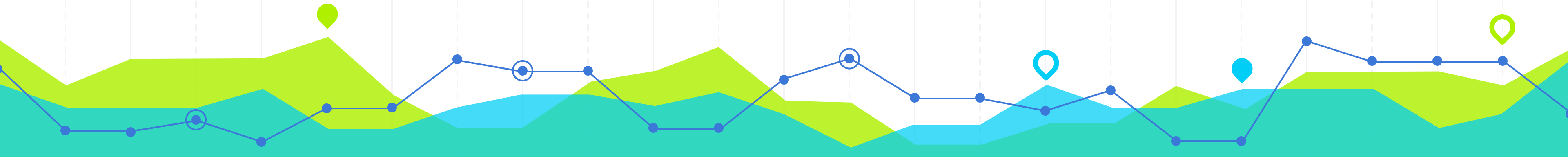
Python

- Faster than R
- Still free
- Very popular
- General programming lang, not specific to data science
  - New methods often released first in R
- Writing code takes longer

Julia

- Fastest
- Also free
- New kid on the block
- Newness means small user base, fewer packages

Solution: Work in multiple languages, R for prototyping, Python for production



Are You Ready To Learn R?

If you say no, I don’t have anything else planned …

7

Get your computers out!

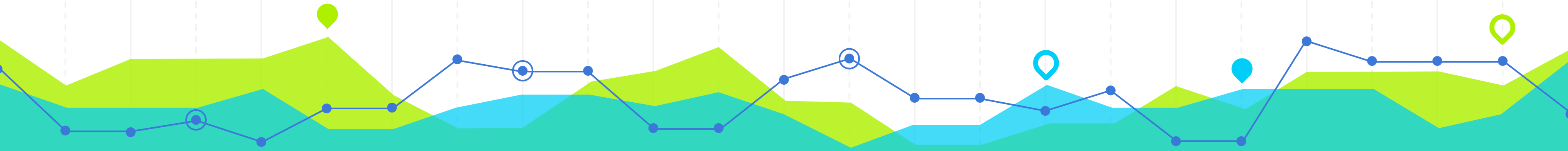
https://www.collinmmccabe.com/fiddle.html

If you haven’t downloaded R,
you can do the exercises
online using an R fiddle

```
inal String GREETING_
tic final String NAME_UNKNOWN =

c static void main(String[] args) {
  if (args.length > 0) {
      greeting(args[0]);
  } else {
      greeting();
  }
}

private static void greeting() {
    greeting(NAME_UNKNOWN);
}

/**
 * Виводить рядок з привітанням (зазвичай, у консоль)
 * @param name ім'я особи, до якої звернене привітан
 /
ate static void greeting(String name) {
stem.out.println(String.format(GREETING_T
```

Code is at:

https://github.com/collinmmccabe/r-workshop

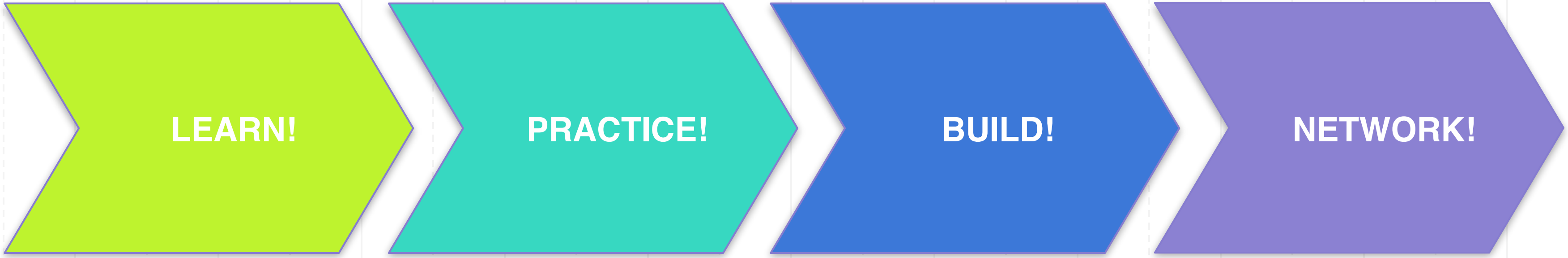


8

Pretty neat, huh?

… but now what do you do?

How can you become a data scientist?



LEARN!

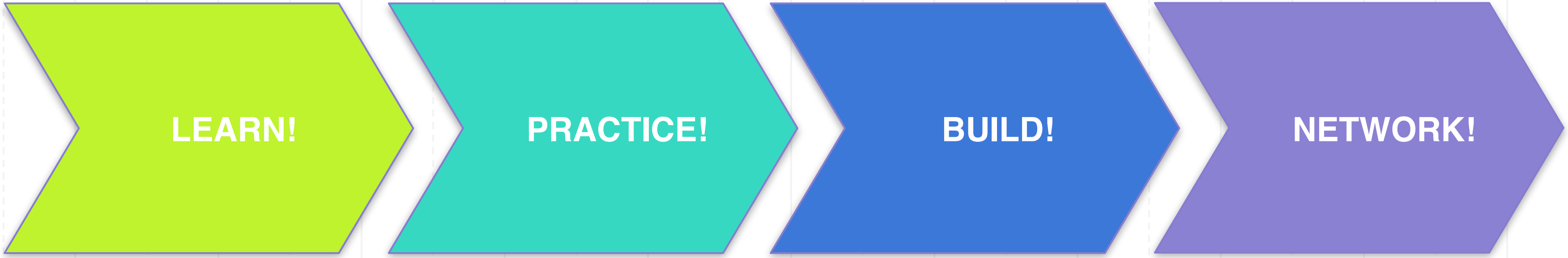
PRACTICE!

BUILD!

NETWORK!

- Bootcamps (I teach one at OSU called Códe with the Erdós Institute)
- Online courses (DataCamp is particularly good for R and Python)
- Online/physical books (R for Data Science, R4DS, is a good one)
- Grow your toolbox: learn Python, SQL, Hadoop, Spark, Tensorflow, etc
- Consider going to grad school (Many data scientists have adv degrees)

How can you become a data scientist?


LEARN!

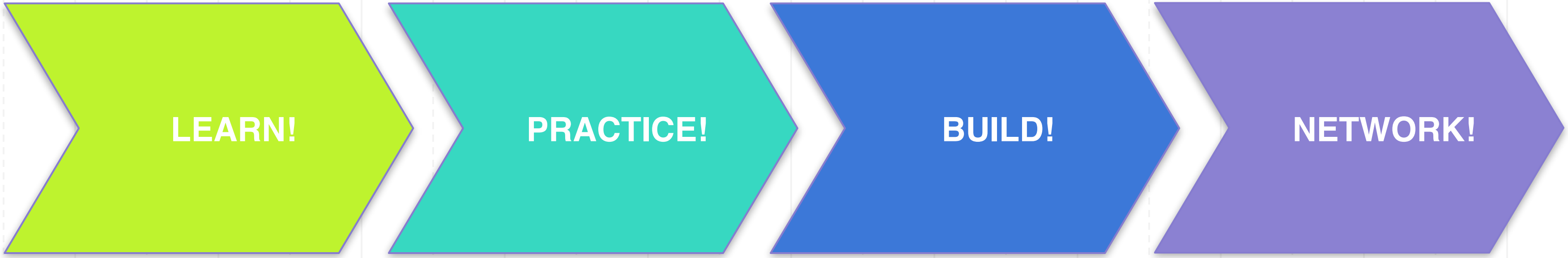
PRACTICE!

BUILD!

NETWORK!

- Kaggle
- DataCamp Projects
- Work through common example datasets (UC Irvine Machine Learning Repository: iris, NMIST, etc)
- Use R for school projects

How can you become a data scientist?



LEARN!

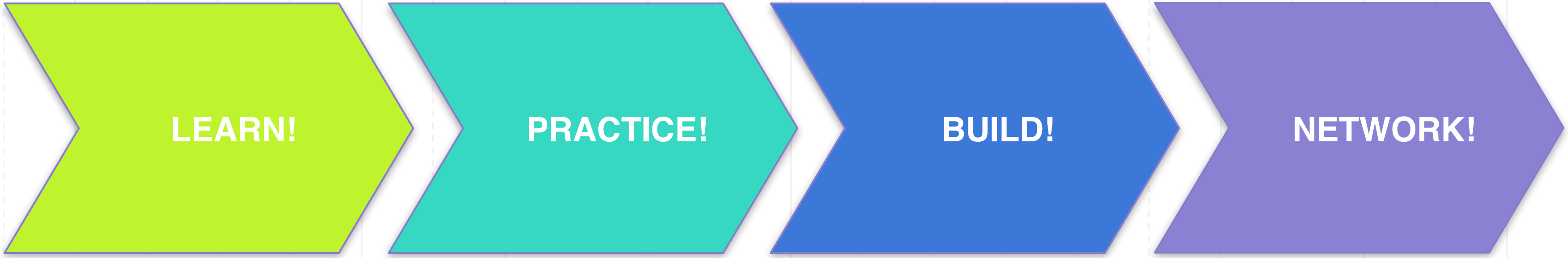
PRACTICE!

BUILD!

NETWORK!

- Side Projects: Focus on your interests
- Build together: Contribute to open-source
- Compete in Hackathons:
    - TechStars Startup Weekend
    - OHI/O (Hack OHI/O, Logi OHI/O, Data OHI/O, others?)

How can you become a data scientist?


LEARN!

PRACTICE!

BUILD!

NETWORK!

- Get involved with data science / analytics groups on campus
- Reach out to data scientists on professional and social networks
- Talk to your professors, ask if they know people in data science
- Use university resources like Buckeye Careers, CCSS for connecting with companies

THANKS!

Any questions?

You can find me at

@collinmmccabe / collin.michael.mccabe@gmail.com

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- R created and maintained by The R Foundation
- Presentation template by SlidesCarnival
- Images by Wikimedia Commons