Practical Data Science I:
Wrangling Data and Answering Questions

Adriane Fresh & Nick Eubank

What is Data Science?

What is Data Science?

1. What (in theory) do we think Data Science should be?

What is Data Science?

1. What (in theory) do we think Data Science should be?
2. What (empirically) is Data Science?

What (in theory) should Data Science be?

What (in theory) should Data Science be?

Study of how best to answer questions using quantitative data.

What (empirically) is Data Science?

How did Data Science become a thing?

Over the past several decades:

1. Availability of data ↑
2. Computational power ↑

How did Data Science become a thing?

Over the past several decades:

1. Availability of data $\uparrow$
2. Computational power $\uparrow$

$\Rightarrow$ Huge proliferation and increase in sophistication of computational methods

How did Data Science become a thing?

- Academic research is organized into silos:

How did Data Science become a thing?

- Academic research is organized into silos:
    - Computer Science
    - Statistics
    - Economics
    - Political science
    - Engineering

How did Data Science become a thing?

- Academic research is organized into silos:
    - Computer Science
    - Statistics
    - Economics
    - Political science
    - Engineering

⇒ Development of new tools occurred *within* each silo.

Where are we today?

Very little cross-pollination across silos

Where are we today?

Very little cross-pollination across silos

- Lots of duplication of development.

Where are we today?

Very little cross-pollination across silos

- Lots of duplication of development.
- Every silo has its own vocabulary.

Where are we today?

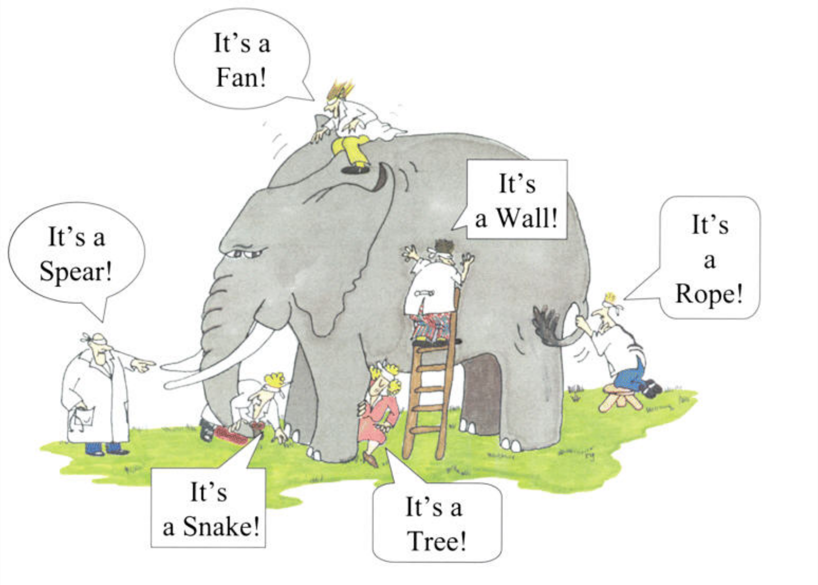
Very little cross-pollination across silos

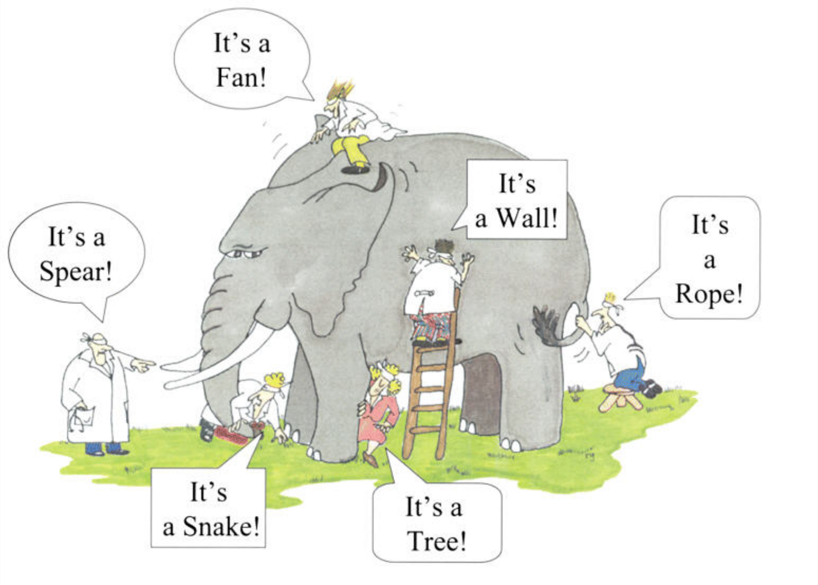
- Lots of duplication of development.
- Every silo has its own vocabulary.
- Each silo has focused on the aspects most relevant to their applications. e.g.:

Where are we today?

Very little cross-pollination across silos

- Lots of duplication of development.
- Every silo has its own vocabulary.
- Each silo has focused on the aspects most relevant to their applications. e.g.:
    - CS likes to classify things and make predictions, don't care how model works
    - Social scientists like to make causal statements, don't care about predictive power


It's a Fan!
It's a Spear!
It's a Wall!
It's a Rope!
It's a Snake!
It's a Tree!



⇒ This is where we are *now.*

What is (empirically) Data Science?

What is (empirically) Data Science?

An effort to unify the development of quantitative methods

What is (empirically) Data Science?

An effort to unify the development of quantitative methods
→ Recognize the elephant

Why does this matter to you?

- Most current researchers learned their skills in a silos.

Why does this matter to you?

- Most current researchers learned their skills in a silos.
  In many ways, *you* will have better perspective than your professors.

Why does this matter to you?

- Most current researchers learned their skills in a silos.
  In many ways, *you* will have better perspective than your professors.
- Expect fragmentation in both software and vocabulary.

Why does this matter to you?

- Most current researchers learned their skills in a silos.
  In many ways, *you* will have better perspective than your professors.
- Expect fragmentation in both software and vocabulary.
  The fact things don't always seem coherent isn't because you're missing something.

This Class

“The Python Class”

This Class

“The Python Class”

- *Data Science* Python

This Class

"The Python Class"

- *Data Science* Python
    - Python standard library, numpy, pandas, scikit-learn, matplotlib, statsmodels

“The Everything-That-Comes-Before-Data-Analysis Class”

“The Everything-That-Comes-Before-Data-Analysis Class”

In this course, you will learn:

“The Everything-That-Comes-Before-Data-Analysis Class”

In this course, you will learn:

- Think computationally,

“The Everything-That-Comes-Before-Data-Analysis Class”

In this course, you will learn:

- Think computationally,
- Write your own algorithms and generalized code,

“The Everything-That-Comes-Before-Data-Analysis Class”

In this course, you will learn:

- Think computationally,
- Write your own algorithms and generalized code,
- about different data formats and how to work with them,

“The Everything-That-Comes-Before-Data-Analysis Class”

In this course, you will learn:

- Think computationally,
- Write your own algorithms and generalized code,
- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,

“The Everything-That-Comes-Before-Data-Analysis Class”

In this course, you will learn:

- Think computationally,
- Write your own algorithms and generalized code,
- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,
- best practices for data science workflow management and collaboration.

“The Everything-That-Comes-Before-Data-Analysis Class”

In this course, you will learn:

- Think computationally,
- Write your own algorithms and generalized code,
- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,
- best practices for data science workflow management and collaboration.

All through hands-on experience.

This Class

So yes, we will be working *in* Python,

This Class

So yes, we will be working *in* Python,
but this isn't a class *about* Python.

This Class

So yes, we will be working *in* Python,
but this isn't a class *about* Python.
⇒ Emphasis on generalizable data science skills

This Class

By the end of this class:

This Class

By the end of this class:

- Find and organize data *on your own*.

This Class

By the end of this class:

- Find and organize data *on your own*.
- Understand how to clean, merge, and manipulate real-world data.

This Class

By the end of this class:

- Find and organize data *on your own*.
- Understand how to clean, merge, and manipulate real-world data.
- Know how to approach organizing a full project.

Why Python?

Why Python?

And why not:

- Stata
- R (Tidy-Verse)
- R (Base-R)

?

Why Python?

And why not:

- Stata

  Excellent for tabular data, some text
- R (Tidy-Verse)
- R (Base-R)

?

Why Python?

And why not:

- Stata

  Excellent for tabular data, some text
- R (Tidy-Verse)

  Excellent for tabular data
- R (Base-R)

  Good for tabular, network, geospatial, some text and ML

?

Why Python?

Python:

- Tabular data
- Network data
- Geospatial data
- Natural Language Processing (NLP)
- Neural Networks
- Using with Cloud Compute
- Big Data
- Large Language Models
- All Machine Learning
- ...

Why?

Language Intrinsics:

Why?

Language Intrinsics:

- R and Stata are *domain-specific languages* (DSLs).

Why?

Language Intrinsics:

- R and Stata are *domain-specific languages* (DSLs).
  - Simplified to make them easier for researchers to start using.

Why?

Language Intrinsics:

- R and Stata are *domain-specific languages* (DSLs).
    - Simplified to make them easier for researchers to start using.
- Python is a *general purpose language*.

Why?

Language Intrinsics:

- R and Stata are *domain-specific languages* (DSLs).
    - Simplified to make them easier for researchers to start using.
- Python is a *general purpose language*.
    - Python is foundational at OpenAI, Instagram, Dropbox, Netflix.

Why?

Language Intrinsics:

- R and Stata are *domain-specific languages* (DSLs).
  - Simplified to make them easier for researchers to start using.
- Python is a *general purpose language*.
  - Python is foundational at OpenAI, Instagram, Dropbox, Netflix.

Network Effects:

Why?

Language Intrinsics:

- R and Stata are *domain-specific languages* (DSLs).
  - Simplified to make them easier for researchers to start using.
- Python is a *general purpose language*.
  - Python is foundational at OpenAI, Instagram, Dropbox, Netflix.

Network Effects:

- 90s and 2000s (even 2010s): most social scientists used Stata or R, developed for Stata and R.

Why?

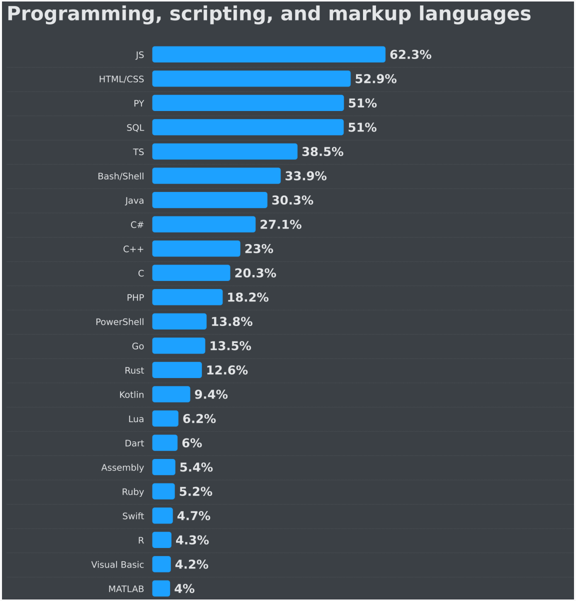
Language Intrinsics:

- R and Stata are *domain-specific languages* (DSLs).
  - Simplified to make them easier for researchers to start using.
- Python is a *general purpose language*.
  - Python is foundational at OpenAI, Instagram, Dropbox, Netflix.

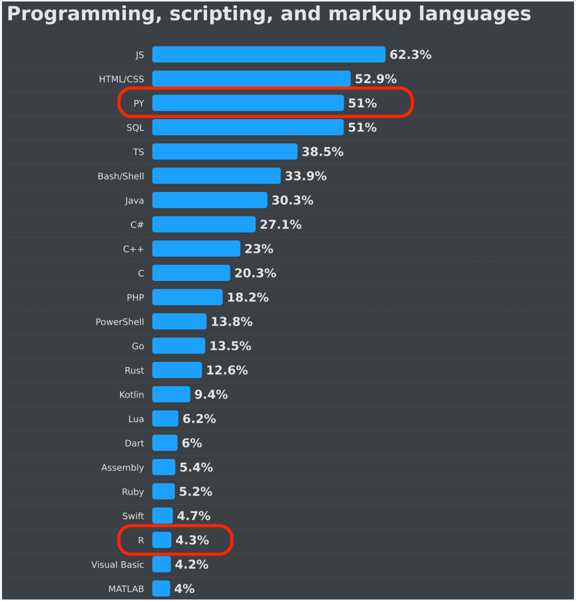
Network Effects:

- 90s and 2000s (even 2010s): most social scientists used Stata or R, developed for Stata and R.
- 2010s and 2020s: Businesses, computer scientists and software engineers moved to data science.
  - Wanted a fully-featured, general purpose language (ideally one they know).

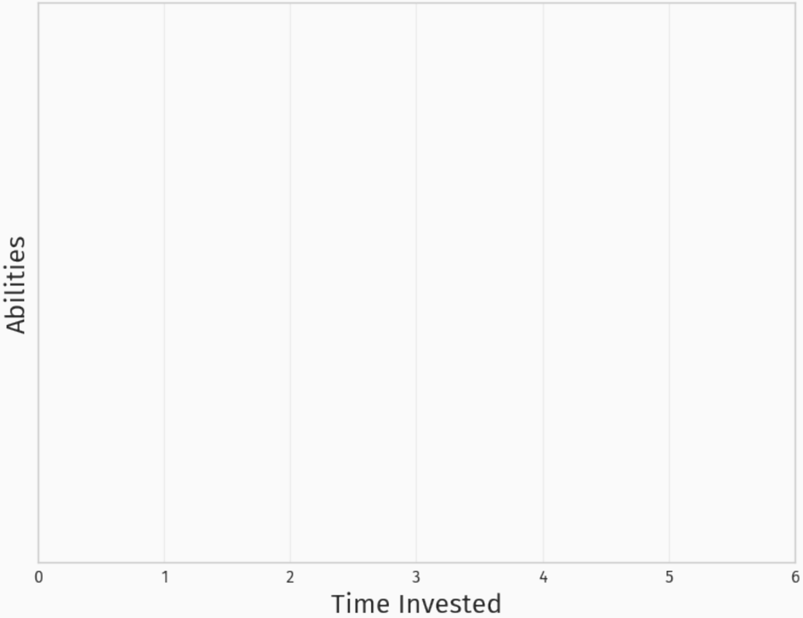
Stack Overflow Developer Survey 2024

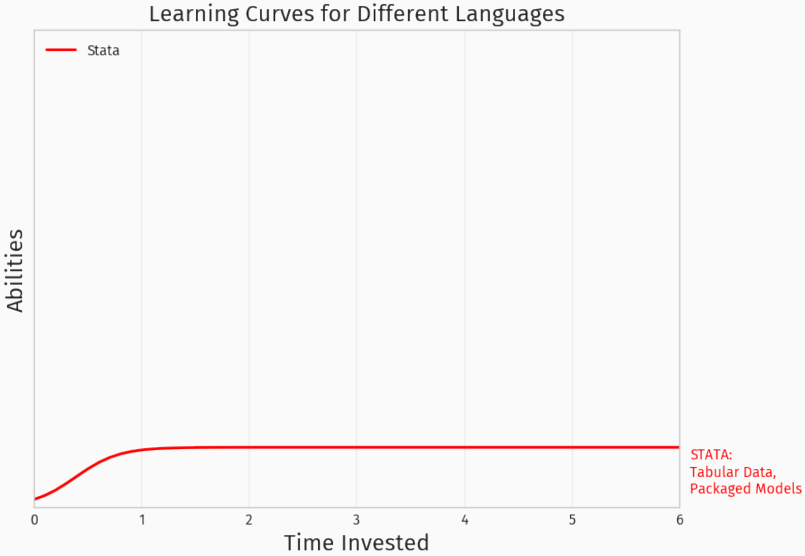


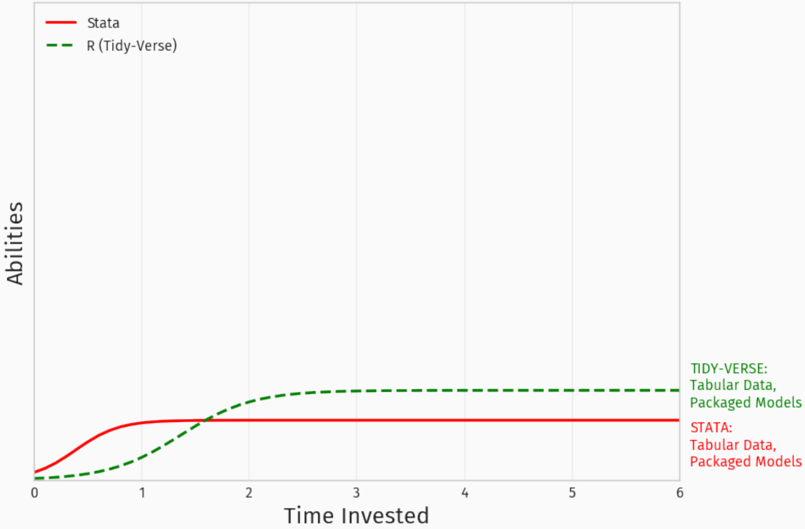
Stack Overflow Developer Survey 2024



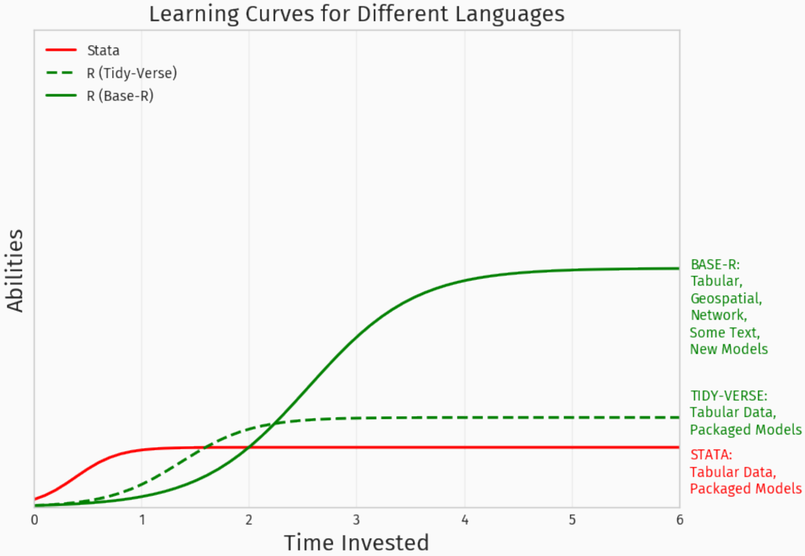
Learning Curves for Different Languages

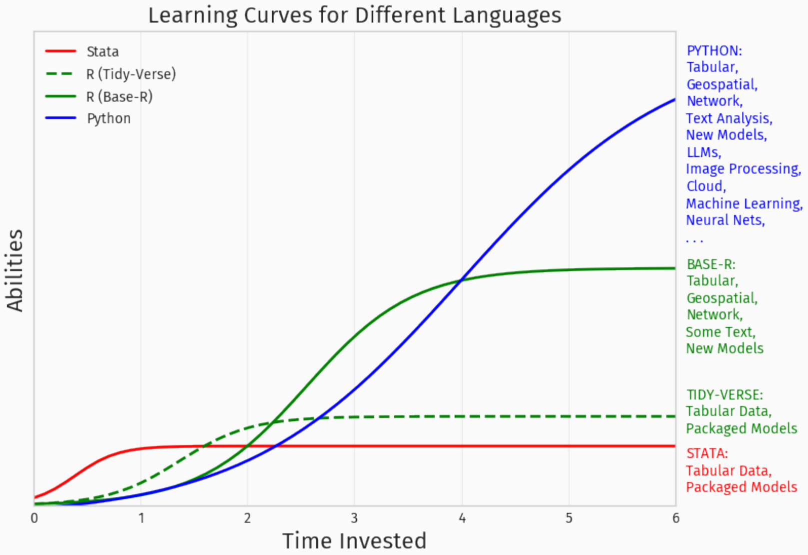



Learning Curves for Different Languages
Stata
Abilities
STATA:
Tabular Data,
Packaged Models
0
1
2
3
4
5
6
Time Invested



Learning Curves for Different Languages


Learning Curves for Different Languages
Stata
R (Tidy-Verse)
R (Base-R)
Abilities
Time Invested
0
1
2
3
4
5
6
BASE-R:
Tabular,
Geospatial,
Network,
Some Text,
New Models
TIDY-VERSE:
Tabular Data,
Packaged Models
STATA:
Tabular Data,
Packaged Models



Learning Curves for Different Languages

So what?

So what?

- “This is so much easier to do in [Stata/R]”
  - You’re not wrong!

So what?

- “This is so much easier to do in [Stata/R]”
  - You’re not wrong!
  - (Also easier ways in Python!)

So what?

- “This is so much easier to do in [Stata/R]”
  - You’re not wrong!
  - (Also easier ways in Python!)
- “This isn’t what I want to learn.”
  - Especially for the first 4 weeks.

By the end of this class...

- Manipulating and cleaning US census data,

By the end of this class...

- Manipulating and cleaning US census data,
- Reshaping and aggregating arrest data,

By the end of this class...

- Manipulating and cleaning US census data,
- Reshaping and aggregating arrest data,
- Statistically estimating the effect of smoking on infant birthweight,

By the end of this class...

- Manipulating and cleaning US census data,
- Reshaping and aggregating arrest data,
- Statistically estimating the effect of smoking on infant birthweight,
- and more.

If you also take IDS 591...

- Work with big (terabyte sized) datasets,

If you also take IDS 591...

- Work with big (terabyte sized) datasets,
- Analyze geospatial satellite and demographic data,

If you also take IDS 591...

- Work with big (terabyte sized) datasets,
- Analyze geospatial satellite and demographic data,
- Analyze social networks,

If you also take IDS 591...

- Work with big (terabyte sized) datasets,
- Analyze geospatial satellite and demographic data,
- Analyze social networks,
- and more.

Most importantly, though...

You will have a strong understanding of computational thinking and Python

Most importantly, though...

You will have a strong understanding of computational thinking and Python
that will allow your empirical work to go wherever your research takes you,

Most importantly, though...

You will have a strong understanding of computational thinking and Python
that will allow your empirical work to go wherever your research takes you,
instead of feeling limited by what existing packages make easy.

About Us: Adriane Fresh

I am a scholar of political economy and political institutions

About Us: Adriane Fresh

I am a scholar of political economy and political institutions

- PhD in Political Economy

About Us: Adriane Fresh

I am a scholar of political economy and political institutions

- PhD in Political Economy
- Master in Economics

About Us: Adriane Fresh

I am a scholar of political economy and political institutions

- PhD in Political Economy
- Master in Economics
- BA in Economics and Latin American Studies, Minor in Math

About Us: Adriane Fresh

Research:

About Us: Adriane Fresh

Research:

- Effects of economic and institutional changes on elite persistence and the strategies that elites.

About Us: Adriane Fresh

Research:

- Effects of economic and institutional changes on elite persistence and the strategies that elites.
- I study a diverse set of historical time periods and country contexts including the Industrial Revolution in Britain, the enfranchisement of Black people in the U.S., regime change in Chile, and contemporary U.S. election administration.

About Us: Adriane Fresh

Research:

- Effects of economic and institutional changes on elite persistence and the strategies that elites.
- I study a diverse set of historical time periods and country contexts including the Industrial Revolution in Britain, the enfranchisement of Black people in the U.S., regime change in Chile, and contemporary U.S. election administration. I am particularly interested in causal inference and natural language processing using large corpuses of historical and historiographical text.

About Us: Nick Eubank

I am en empirical social scientist

About Us: Nick Eubank

I am en empirical social scientist

- PhD in Political Science

About Us: Nick Eubank

I am en empirical social scientist

- PhD in Political Science
- Master in Economics

About Us: Nick Eubank

I am en empirical social scientist

- PhD in Political Science
- Master in Economics
- BA in Economics and Political Science

About Us: Nick Eubank

Research:

About Us: Nick Eubank

Research:

- Looking for evidence of polling place manipulation in North Carolina

About Us: Nick Eubank

Research:

- Looking for evidence of polling place manipulation in North Carolina
- Developing methods of measuring Gerrymandering in the US.

About Us: Nick Eubank

Research:

- Looking for evidence of polling place manipulation in North Carolina
- Developing methods of measuring Gerrymandering in the US.
- Testing theories about how social networks shape political behavior using cell-phone meta-data to map social networks of entire countries (Zambia and Venezuela).

About Us: Nick Eubank

Research:

- Looking for evidence of polling place manipulation in North Carolina
- Developing methods of measuring Gerrymandering in the US.
- Testing theories about how social networks shape political behavior using cell-phone meta-data to map social networks of entire countries (Zambia and Venezuela).
- Studying whether political elites in the US South turned to using incarceration to prevent black voters from exercising political influence after the Voting Rights Act removed their ability to use Jim Crow restrictions.

Features of this class

- Flipped Classroom

Features of this class

- Flipped Classroom
    - Need to come prepared,

Features of this class

- Flipped Classroom
  - Need to come prepared,
  - Intermittent Reading Quizzes

Features of this class

- Flipped Classroom
    - Need to come prepared,
    - Intermittent Reading Quizzes
- Lots of group work

Features of this class

- Flipped Classroom
  - Need to come prepared,
  - Intermittent Reading Quizzes
- Lots of group work
  - Pair programming

Features of this class

- Flipped Classroom
  - Need to come prepared,
  - Intermittent Reading Quizzes
- Lots of group work
  - Pair programming
  - You'll provide feedback to your partners

Features of this class

- Flipped Classroom
  - Need to come prepared,
  - Intermittent Reading Quizzes
- Lots of group work
  - Pair programming
  - You'll provide feedback to your partners
  - Don't start early

Features of this class

- Flipped Classroom
  - Need to come prepared,
  - Intermittent Reading Quizzes
- Lots of group work
  - Pair programming
  - You'll provide feedback to your partners
  - Don't start early
- Slack

Features of this class

- Flipped Classroom
  - Need to come prepared,
  - Intermittent Reading Quizzes
- Lots of group work
  - Pair programming
  - You'll provide feedback to your partners
  - Don't start early
- Slack
- Datacamp

Features of this class

- Flipped Classroom
    - Need to come prepared,
    - Intermittent Reading Quizzes
- Lots of group work
    - Pair programming
    - You'll provide feedback to your partners
    - Don't start early
- Slack
- Datacamp
- Photos

www.practicaldatascience.org