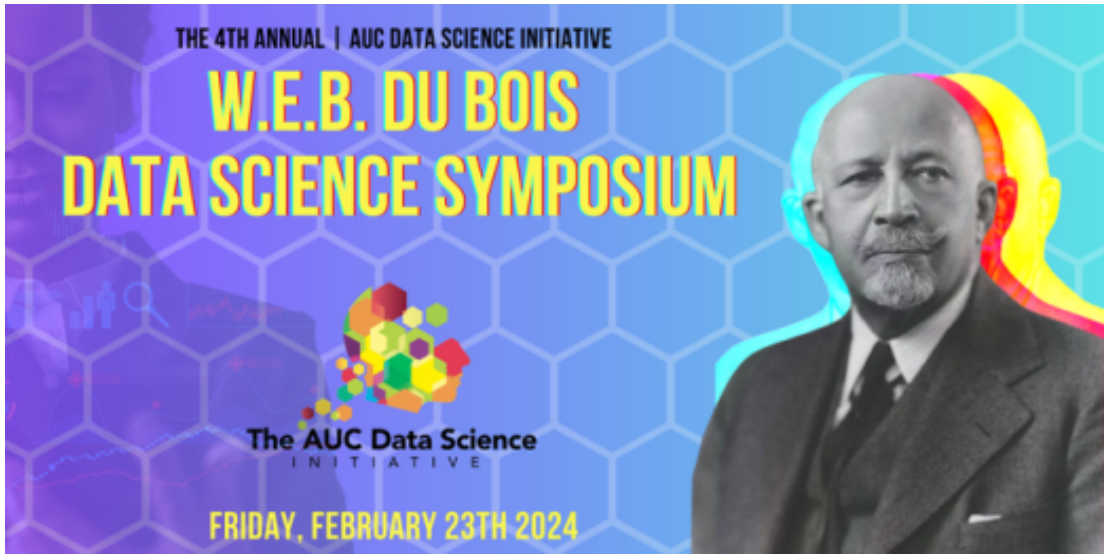


Infusion of Data Science and Computation into Introductory Statistics

Sayed Mostafa

Department of Mathematics & Statistics
North Carolina A&T State University

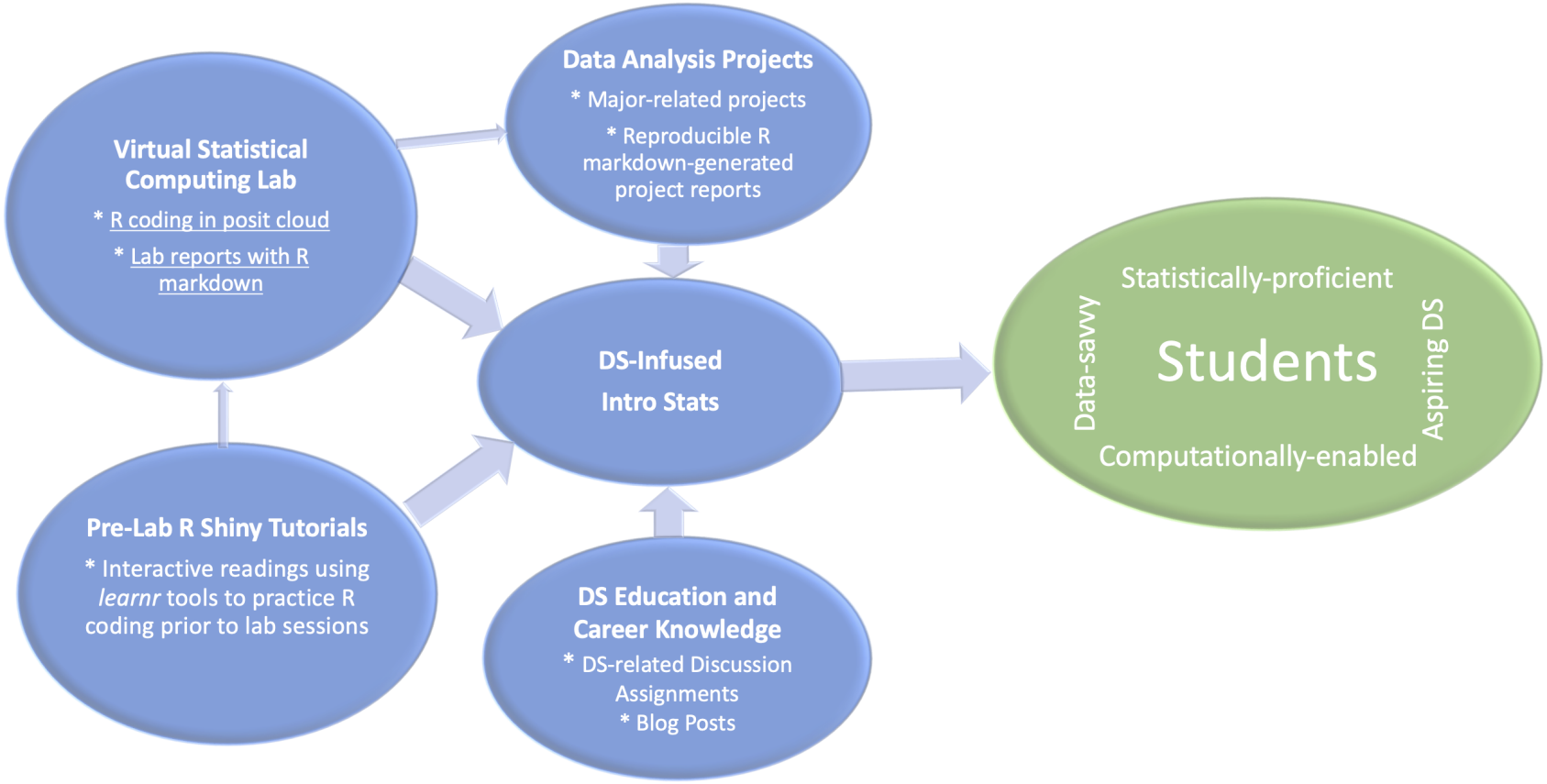
Outline

- Why bring Data Science (DS) and Computation to Introductory Statistics?
- A Proposed Design for a DS/Computationally-Infused Intro Stats Course
- Evaluation Results of the Proposed Design for Intro Stats
- Resources for Teaching a DS-Infused Intro Stats Course

Why introduce DS/computation in Intro Stats?

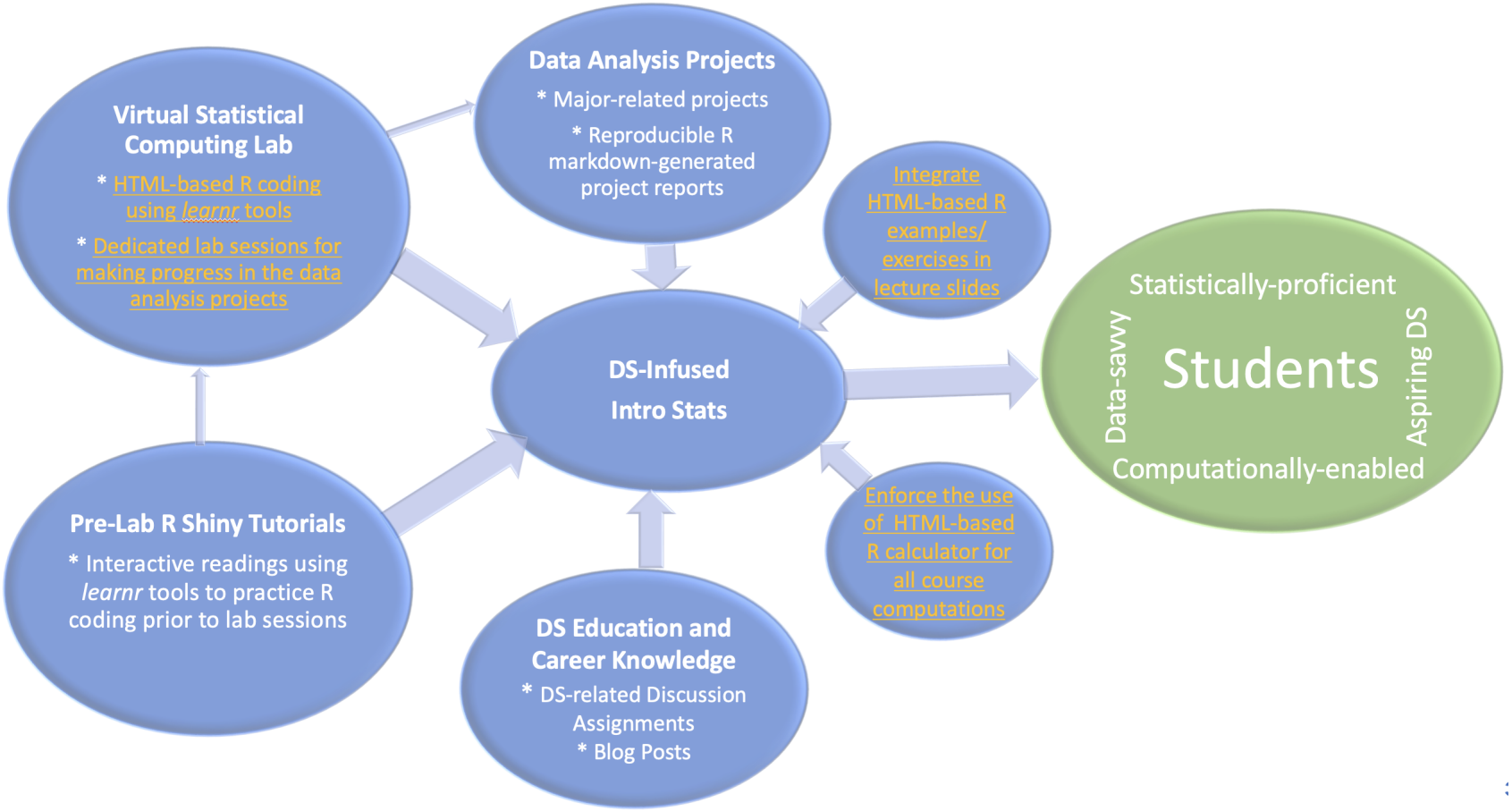
- Fast-growing demand on graduates with computational and data-analytical skills.
- Help all students develop "computational thinking" skills.
- Intro Stats can help us attract and prepare a large diverse pool of UGs for DS education/careers:
  - At NCA&T, Intro Stats is an Algebra-based 3.00 credits course
  - **Large**: 7 sections each semester (~45 students in each section)
  - **Diverse**: serves STEM (~46%) and non-STEM (~54%) majors
- Intro Stats students are likely unaware of DS opportunities:
  - A survey of NCA&T's Intro Stats students ($n = 181$) found that
    - Only **33.15%** of students surveyed had heard about DS,
    - Of those, only **27.12%** knew NCA&T offers DS courses.

DS/Computationally-Infused Intro Stats: Phase I-FA22



- **Implementation**: 2 treatment sections and 2 control sections

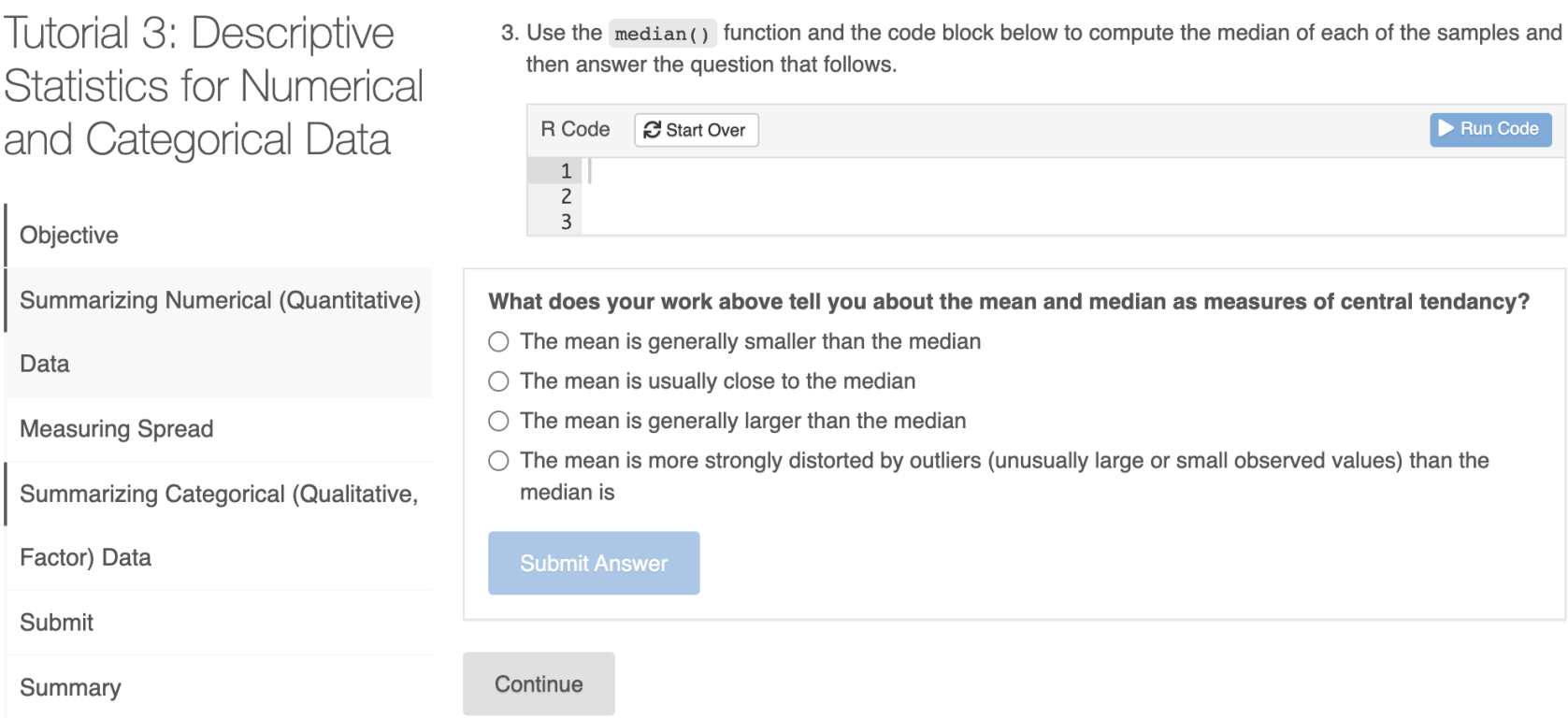
DS/Computationally-Infused Intro Stats: Phase II-SP23



- **Implementation**: 4 treatment sections and 2 control sections

DS/Computationally-Infused Intro Stats: Phase II-SP23

- Interactive Shiny Pre-Lab Tutorial



DS/Computationally-Infused Intro Stats: Phase II-SP23

- Interactive Computing Lab (using the *learnr* package)

Exploratory Data Analysis Part I

Start Over

Recall that the five number summary includes the min, first quantile (Q1), median, third quantile (Q3), and max. Using the `mpg` dataset, we can compute the five number summary of the vehicle's highway mileage `hwy` as follows.

R Code | Start Over | Run Code

```
mpg %>%
  summarize(Min = min(hwy),
            Q1 = quantile(hwy, 0.25),
            Median = median(hwy),
            Q3 = quantile(hwy, 0.75),
            Max = max(hwy)
            )
```

Notice how the `quantile()` function is used to obtain quantiles by setting the proportion of data below the quantile (i.e., 0.25 or 0.75)

4. Use the code chunk below to calculate the measures of center (mean and median) for the vehicle's city mileage `cty`.

R Code | Start Over | Run Code | Submit Answer

5. Use the code chunk below to calculate the variation measures (standard deviation and interquartile range) for the vehicle's city mileage `cty`.

R Code | Start Over | Run Code | Submit Answer

DS/Computationally-Infused Intro Stats: Phase II-SP23

- Interactive R Calculator

Using R as a calculator

R can be used as an calculator as we already saw in the tutorial. So let's get a refresher on this.

Let's say we want to calculate $\frac{36}{29(15-9)}$. Then we would do the following:

R Code | Start Over | Run Code

```
36 / (29 * (15 - 9))


```

R also has built-in constants such as pi and mathematical functions such as e and log.

Let's find the radius of a circle with radius 4. Then using R we can get the area and the circumference.

R Code | Start Over | Run Code

```
radius = 4

area = pi * radius^2

circumference = 2 * pi * radius

c("Area" = area, "Circumference" = circumference)
```

We can also use `R` to calculate probabilities under the normal distribution. The following code returns the probability that a normal variable with mean 25 and standard deviation 15 is less than 50.

R Code | Start Over | Run Code

```
pnorm(q = 50, mean = 25, sd = 15)


```

As you work on your homework assignments, feel free to use the below code chunks to perfom your calculations.

R Code | Start Over | Run Code

```


```

Evaluating the DS-Infused Intro Stats Design

- **DS awareness, readiness & aspirations**
  - Students completed a DS awareness, readiness, and aspirations survey in Qualtrics
  - Pre-survey during 1st week of semester; post-survey at the end of semester
  - The survey was created in-house and validated through a series of exploratory factor analyses using pilot data from SP22
  - Three subscales emerged:
    - Awareness scale (3 items)
    - Readiness scale (4 items)
    - Aspirations scale (5 items)

Awareness of Data Science

- Response Var.: Gain in DS Awareness
- Main Explanatory Var.: Course design (Ref = "Traditional")

Regression Term	Estimate	LCL	UCL	p.value	Sig.
Intercept	0.43	-0.13	0.99	0.1301	Not Sig.
Design: DS-Infused-FA22	0.26	0.08	0.44	0.0050	**
Design: DS-Infused-SP23	0.18	0.02	0.33	0.0230	$*$
Sex: Male	-0.09	-0.24	0.06	0.2439	Not Sig.
Race: Not Black	-0.14	-0.32	0.04	0.1225	Not Sig.
PELL Recepient: Yes	-0.21	-0.40	-0.02	0.0290	$*$
Rural: Yes	0.11	-0.10	0.32	0.3135	Not Sig.
Residency: Out-of-State	0.05	-0.11	0.21	0.5109	Not Sig.
STEM: Yes	-0.15	-0.29	0.00	0.0431	$*$
AP Stat: Yes	0.09	-0.07	0.25	0.2813	Not Sig.
Pre-Course Cum GPA	0.00	-0.14	0.14	0.9858	Not Sig.
Attendance	0.00	0.00	0.01	0.6380	Not Sig.

Significance codes: "$*$" $\rightarrow$ $p.value < 0.05$, "$**$" $\rightarrow$ $p < 0.01$, "$***$" $\rightarrow$ $p < 0.001$, "$****$" $\rightarrow$ $p < 0.0001$.

Readiness for Data Science

- Response Var.: Gain in DS Readiness
- Main Explanatory Var.: Course design (Ref = "Traditional")

Regression Term	Estimate	LCL	UCL	p.value	Sig.
Intercept	0.94	-0.53	2.42	0.2087	Not Sig.
Design: DS-Infused-FA22	0.43	-0.04	0.90	0.0740	Not Sig.
Design: DS-Infused-SP23	0.84	0.46	1.22	0.0000	****
Sex: Male	-0.16	-0.54	0.22	0.4079	Not Sig.
Race: Not Black	-0.21	-0.67	0.25	0.3687	Not Sig.
PELL Recepient: Yes	-0.62	-1.11	-0.13	0.0130	$*$
Rural: Yes	-0.58	-1.10	-0.06	0.0296	$*$
Residency: Out-of-State	0.30	-0.09	0.70	0.1300	Not Sig.
STEM: Yes	-0.06	-0.44	0.32	0.7428	Not Sig.
AP Stat: Yes	-0.27	-0.68	0.14	0.1934	Not Sig.
Pre-Course Cum GPA	0.15	-0.19	0.49	0.3932	Not Sig.
Attendance	0.00	-0.01	0.01	0.9119	Not Sig.

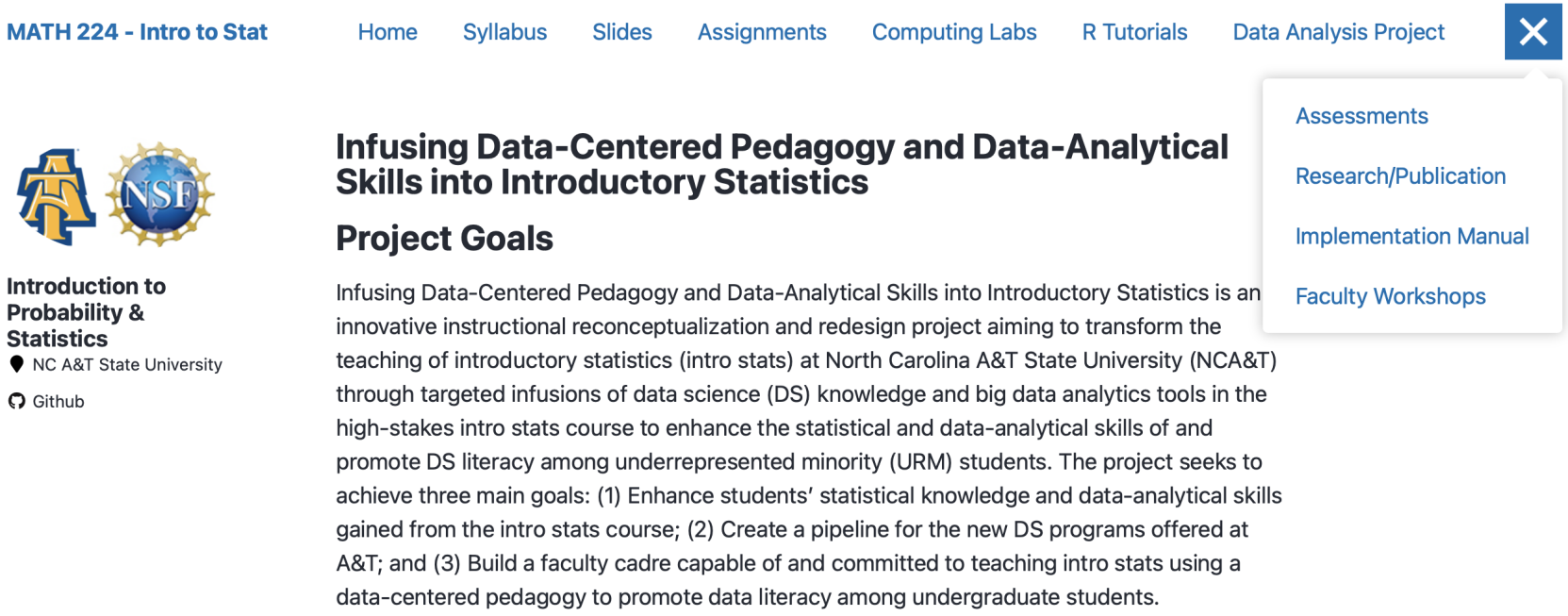
Data Science Aspirations

- Response Var.: Change in DS Aspirations
- Main Explanatory Var.: Course design (Ref = "Traditional")

Regression Term	Estimate	LCL	UCL	p.value	Sig.
Intercept	0.05	-0.53	0.63	0.8688	Not Sig.
Design: DS-Infused-FA22	-0.25	-0.44	-0.07	0.0074	**
Design: DS-Infused-SP23	-0.10	-0.26	0.05	0.2030	Not Sig.
Sex: Male	0.04	-0.11	0.19	0.5946	Not Sig.
Race: Not Black	-0.03	-0.21	0.16	0.7821	Not Sig.
PELL Recepient: Yes	0.14	-0.06	0.33	0.1645	Not Sig.
Rural: Yes	-0.01	-0.22	0.20	0.9514	Not Sig.
Residency: Out-of-State	-0.04	-0.20	0.13	0.6532	Not Sig.
STEM: Yes	-0.04	-0.19	0.11	0.5902	Not Sig.
AP Stat: Yes	0.17	0.01	0.33	0.0426	$*$
Pre-Course Cum GPA	-0.07	-0.22	0.07	0.3156	Not Sig.
Attendance	0.00	0.00	0.01	0.4408	Not Sig.

Resources for Teaching a DS-Infused Intro Stats Course

- Project's Website on GitHub: https://introtostatncat.github.io



- This work is supported by NSF Grant #HRD2106945