Integration of Data Science and Computing into Introductory Statistics

Sayed Mostafa & Tamer Elbayoumi

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

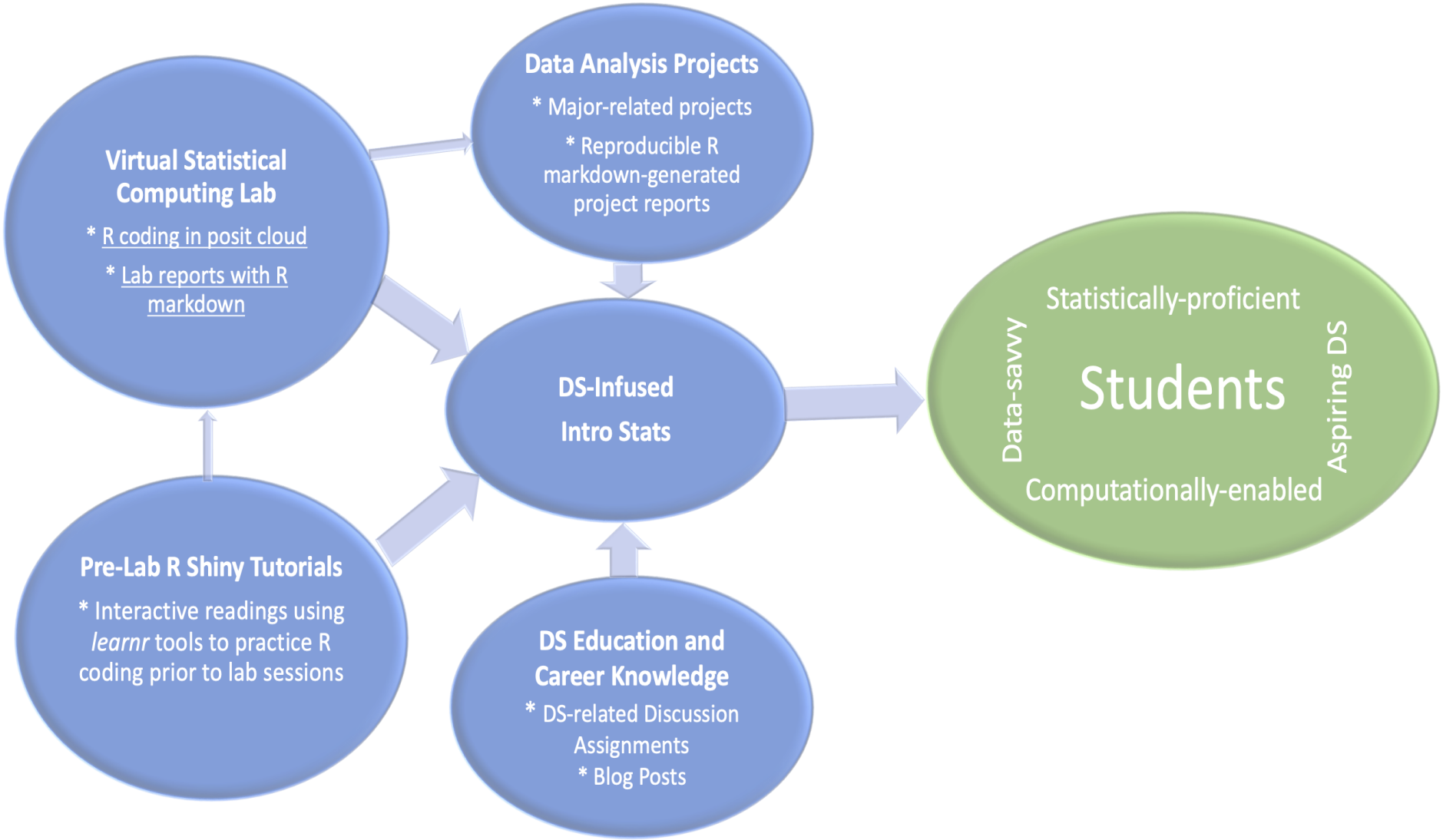
Why introduce DS/computing in Intro Stats?

- 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
- A survey of NCA&T's Intro Stats students ($n = 181$) found that a vast majority are unaware of DS opportunities:
  - Only **33.15%** of students surveyed had heard about DS,
  - Of those, only **27.12%** knew NCA&T offers DS courses.

Guiding Literature

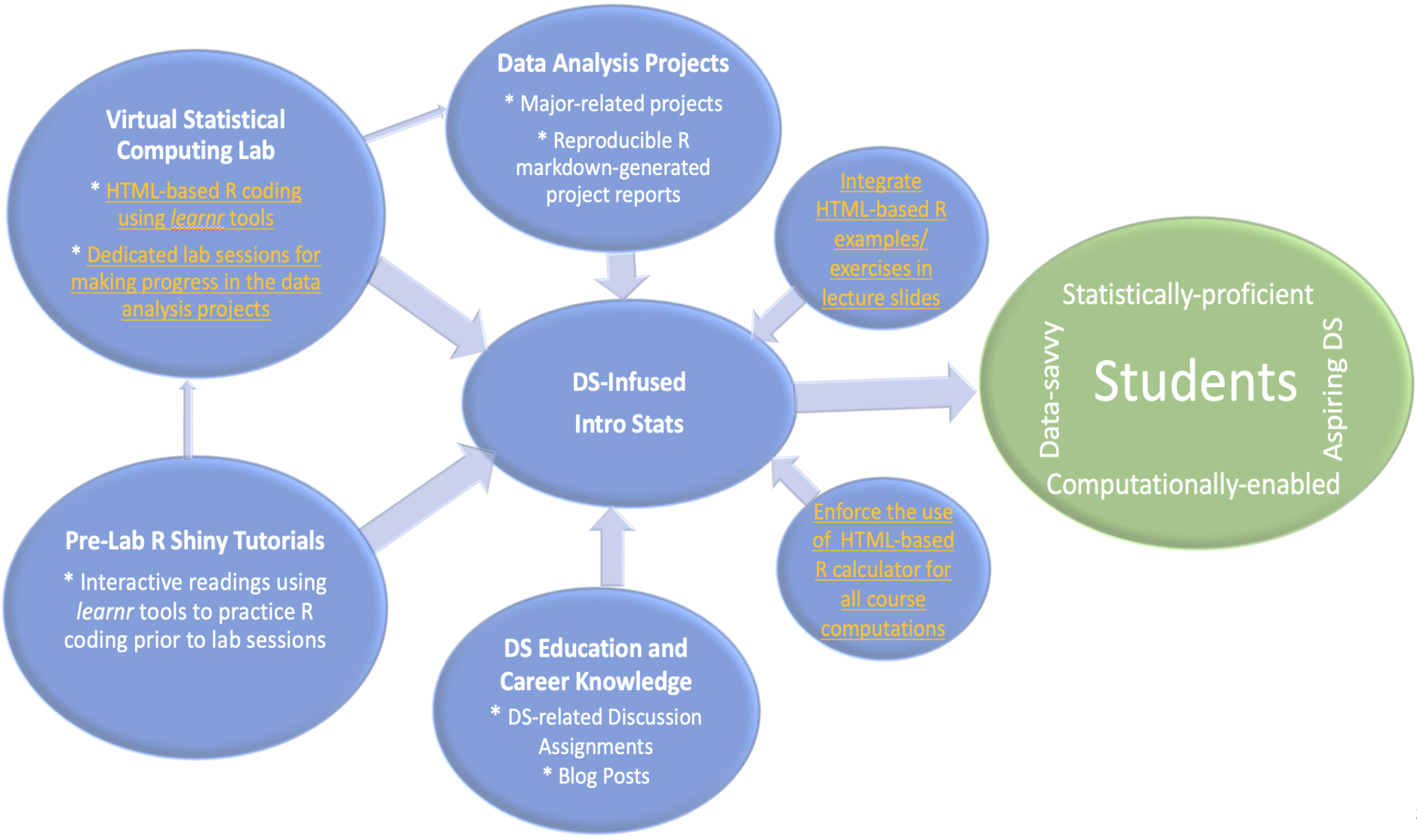
- The Intro Stats course should
  - introduce students to the entire data analysis cycle rather than pieces of it (**Cobb, 2015**),
  - expose students to multivariable thinking (**GAISE #1**),
  - leverage the use of technology for exploring concepts with simulations (**GAISE #2**),
  - help students learn statistics actively while analyzing real data using technology (**GAISE #3, 4 & 5**),
  - train students to think structurally with data and become data-savvy (**Horton et al., 2015**), and
  - expose students, early and frequently, to the elements of the DS workflow and the data scientist's toolbox (**Horton et al., 2015**)
  - See the Special Issue of the *JSDSE* on "**Integrating computing in the statistics and data science curriculum**" (**Horton & Hardin, 2021**).

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

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
- **Statistical learning gains**
  - Students completed a revised version of the CAOS (Comprehensive Assessment of Outcomes in Statistics) scale (e.g., Tintle et al., 2018)
  - Pre/post-test approach

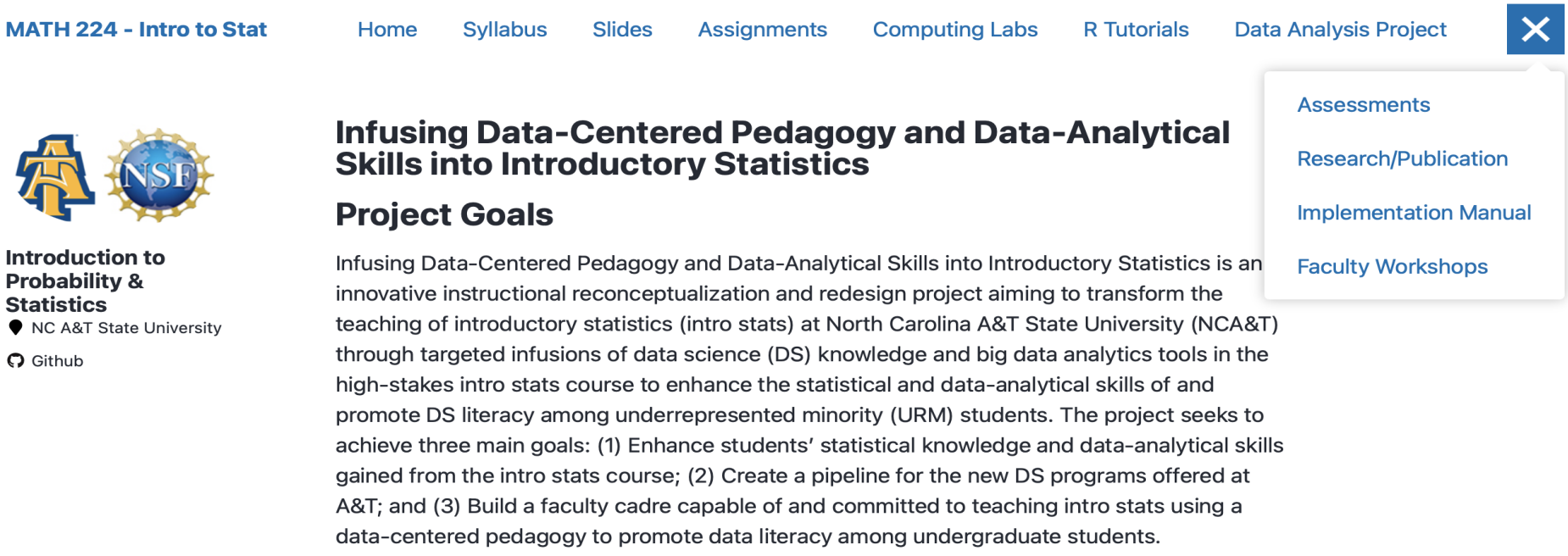
Key Results

Integration of DS tools/knowledge into Intro Stats was associated with

- significant gains in students' levels of DS awareness
  - under both designs (phase I & II)
- significant gains in students' levels of readiness for DS
  - under phase II design only
- significant drop in students' aspirations of DS
  - under phase I design only
- modest statistical learning gains
  - under both designs (phase I & II)

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

- Project Team: Sayed Mostafa; Tamer Elbayoumi; Seongtae Kim; Mingxiang Chen; Guoqing Tang

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.

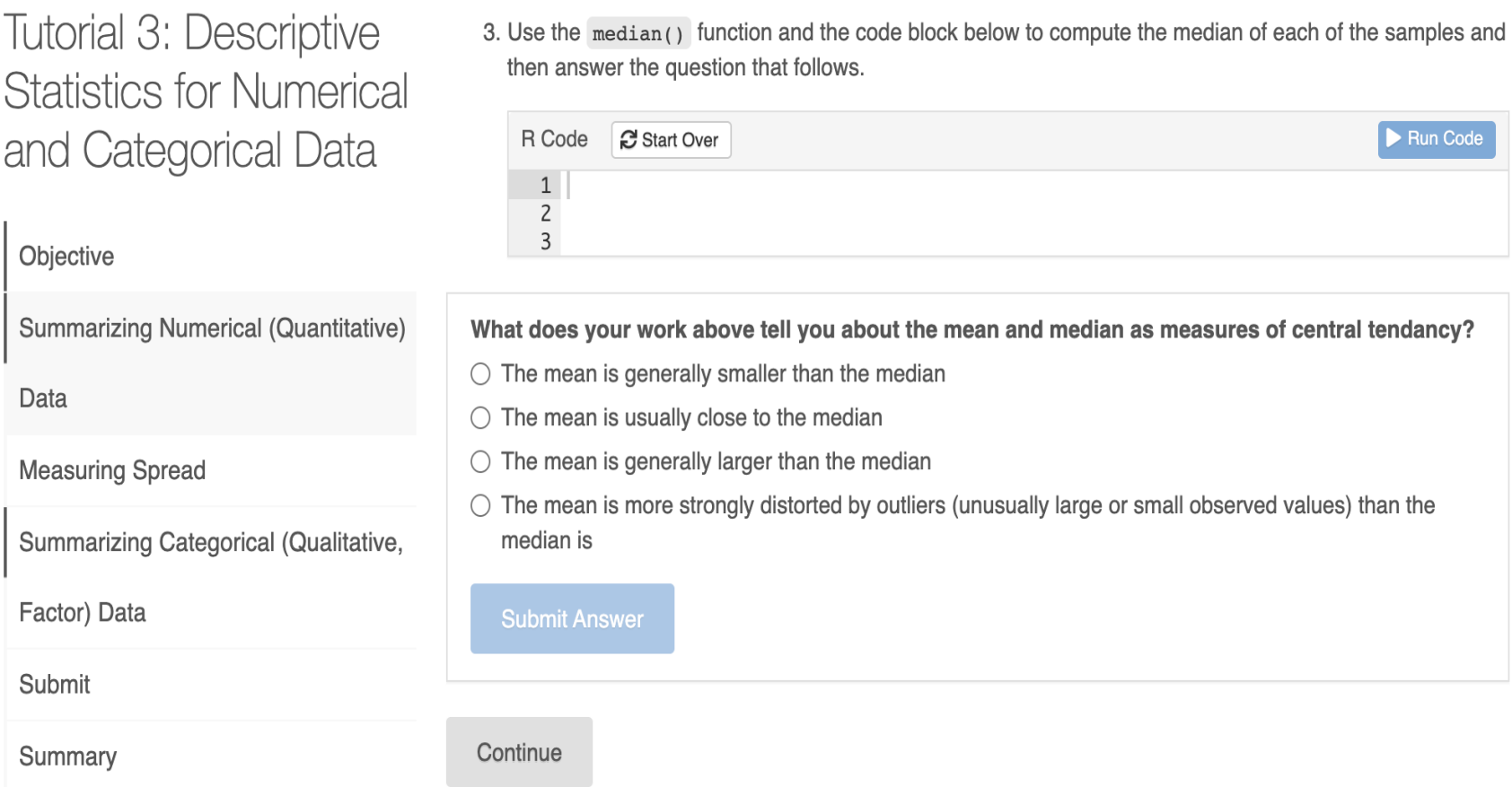
Statistical Learning Gains

- Response Var.: Change in % correct on CAOS test
- Main Explanatory Var.: Course design (Ref = "Traditional")

Regression Term	Estimate	LCL	UCL	p.value	Sig.
Intercept	0.25	-19.37	19.87	0.9800	Not Sig.
Type: DS-Infused-FA22-Design	-0.82	-7.34	5.70	0.8049	Not Sig.
Type: DS-Infused-SP23-Design	0.28	-5.08	5.65	0.9172	Not Sig.
Sex: Male	-1.54	-6.46	3.37	0.5373	Not Sig.
Race: Not Black	1.15	-5.53	7.84	0.7340	Not Sig.
PELL Recepient: Yes	0.69	-5.52	6.89	0.8277	Not Sig.
Rural: Yes	-5.74	-12.38	0.90	0.0901	Not Sig.
Residency: Out-of-State	-6.29	-11.71	-0.88	0.0229	$*$
STEM: Yes	1.14	-3.68	5.95	0.6422	Not Sig.
Pre-Course Cum GPA	-1.96	-6.69	2.76	0.4139	Not Sig.
Attendance	0.18	0.01	0.36	0.0408	$*$

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

- Interactive Shiny Pre-Lab Tutorial (using the *learnr* package)



- Computing Lab Description (Static)

You can also obtain numerical summaries for these flights:

```
lax_flights %>%
  summarise(mean_dd   = mean(dep_delay),
            median_dd = median(dep_delay),
            n         = n())
```

Note that in the `summarise` function you created a list of three different numerical summaries that you were interested in. The names of these elements are user defined, like `mean_dd`, `median_dd`, `n`, and you can customize these names as you like (just don't use spaces in your names). Calculating these summary statistics also requires that you know the function calls. Note that `n()` reports the sample size.

Summary statistics: Some useful function calls for summary statistics for a single numerical variable are as follows:

- `mean()` - The arithmetic mean is found by adding the numbers and dividing the sum by the number of numbers in the list
- `median()` - The middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the mean.
- `sd()` - The measure of the amount of variation or dispersion of a set of values.
- `var()` - the expectation of the squared deviation of a random variable from its population mean or sample mean.
- `IQR()` - the interquartile range is a measure of statistical dispersion, which is the spread of the data. The IQR may also be called the midspread, middle 50%.
- `min()` - The smallest value in the data set.
- `max()` - The largest value in the data set.

Note that each of these functions takes a single vector as an argument and returns a single value.

You can also filter based on multiple criteria. Suppose you are interested in flights headed to San Francisco (SFO) in February:

```
sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)
```

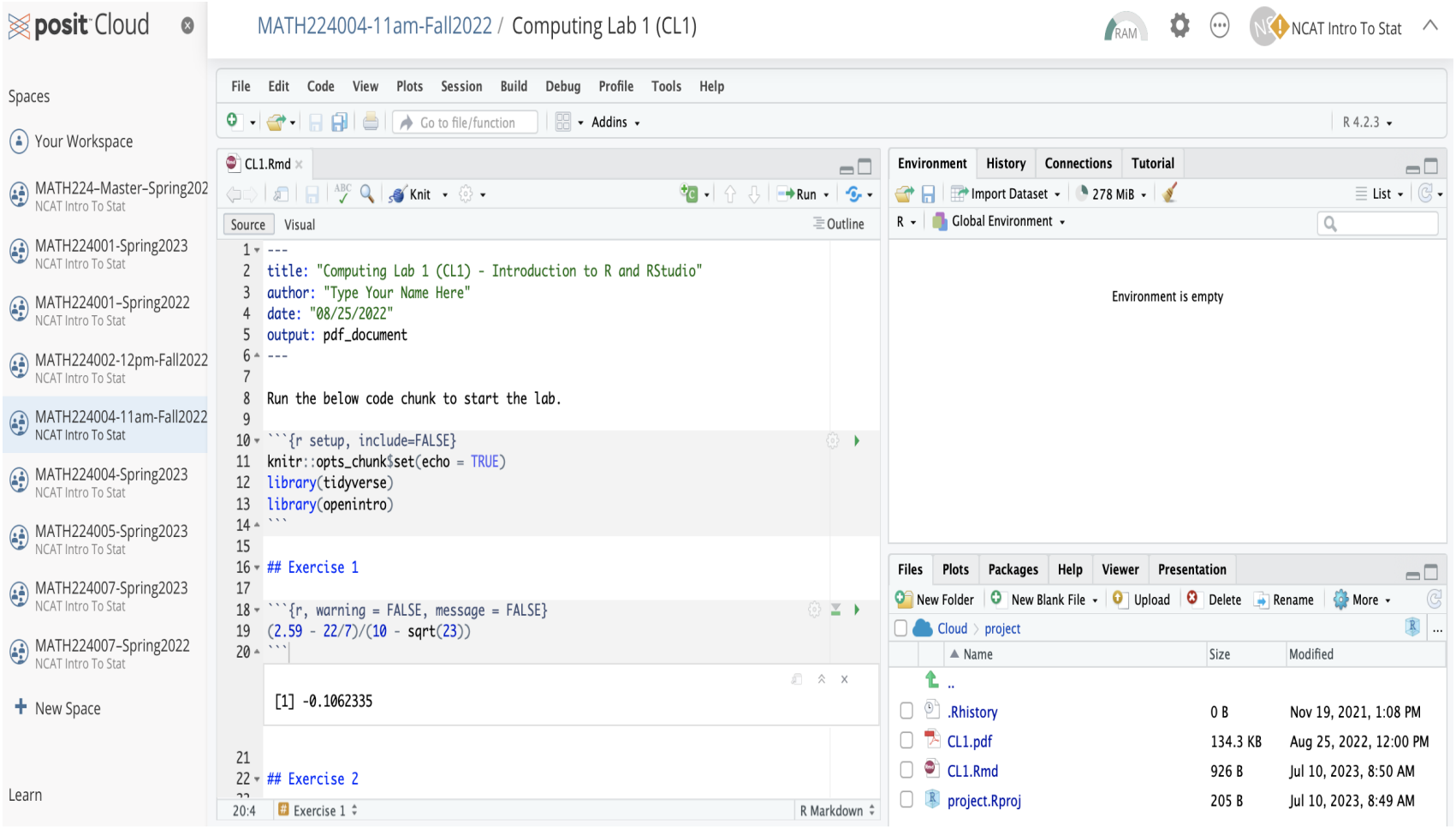
Note that you can separate the conditions using commas if you want flights that are both headed to SFO **and** in February. If you are interested in either flights headed to SFO **or** in February, you can use the `|` instead of the comma.

Exercise 2 Create a new data frame that includes flights headed to SFO in February, and save this data frame as `sfo_feb_flights`. How many flights meet these criteria?

Exercise 3 Describe the distribution of the **arrival** delays of these flights using a histogram and appropriate summary statistics. **Hint:** The summary statistics you use should depend on the shape of the distribution.

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

- Computing Lab R Markdown Template



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

- Slides with Interactive Coding

Examples

Example 1. Calculate the mean of a sample with five observations: 5, 3, 8, 5, 6.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+3+8+5+6}{5} = \frac{27}{5} = 5.4$$

Using R, we can calculate the mean using the `mean()` command. Notice that we need to put the values in a vector using the `c()` function which stands for *concatenate*.

R Code | Start Over | Run Code

```
1 mean(c(5,3,8,5,6))
2
3
```

12/87

Discussions

1. If the data set has 5 observations, with $\bar{x} = 5.4$, find $\sum_{i=1}^{5} x_i$.
2. Continue discussion in 1, if add one more observation 10, will the mean $\bar{x}$ increase or decrease? What is the new $\bar{x}$?
3. Compare data sets 5, 3, 8, 5, 6 and 5, 3, 80, 5, 6, which one has the higher mean?

R Code | Start Over | Run Code

```
1
2
3
```

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

```

```

References I

- Cobb, G. (2015). Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up. *The American Statistician*, 69, 266-282.
- Horton, N.J., Baumer, B.S. and Wickham, H. (2015). Setting the stage for data science: integration of data management skills in introductory and second courses in statistics. *CHANCE*, 28(2):40-50.
- Horton, N.J. and Hardin, J.S. (2021). Integrating computing in the statistics and data science curriculum: Creative structures, novel skills and habits, and ways to teach computational thinking. *Journal of Statistics and Data Science Education*, 29:sup1 S1-S3.
- Tintle, N., Clar, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T. and Vanderstoep, J. (2018). Assessing the Association Between Precourse Metrics of Student Preparation and Student Performance in Introductory Statistics: Results from Early Data on Simulation-Based Inference vs. Nonsimulation-Based Inference. *Journal of Statistics Education*, 26(2), 103-109.
- Woodard, V. and Lee, H. (2021). How students use statistical computing in problem solving. *Journal of Statistics and Data Science Education* 29(1), 1– 18.