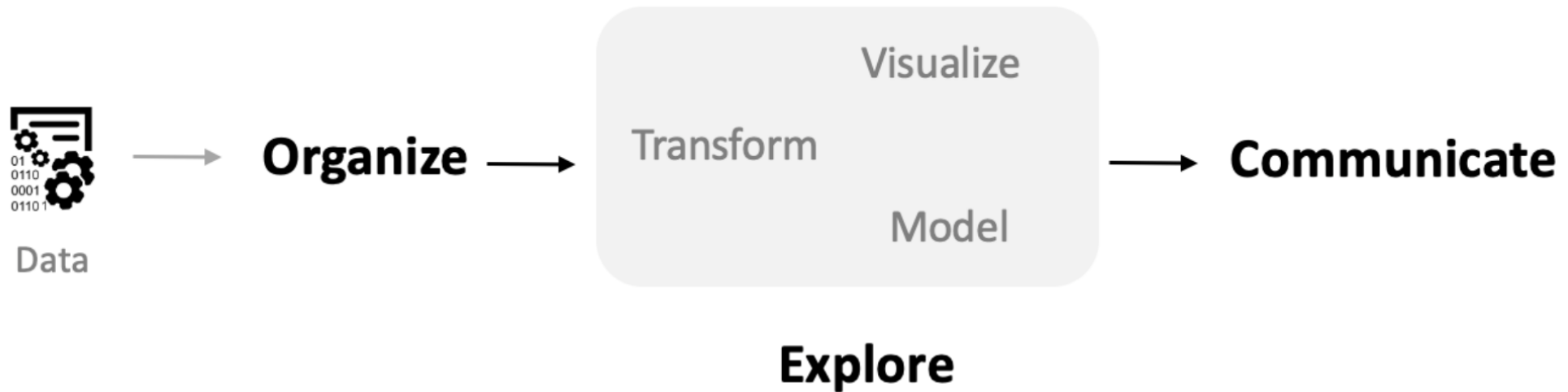


# DSST389: Statistical Learning

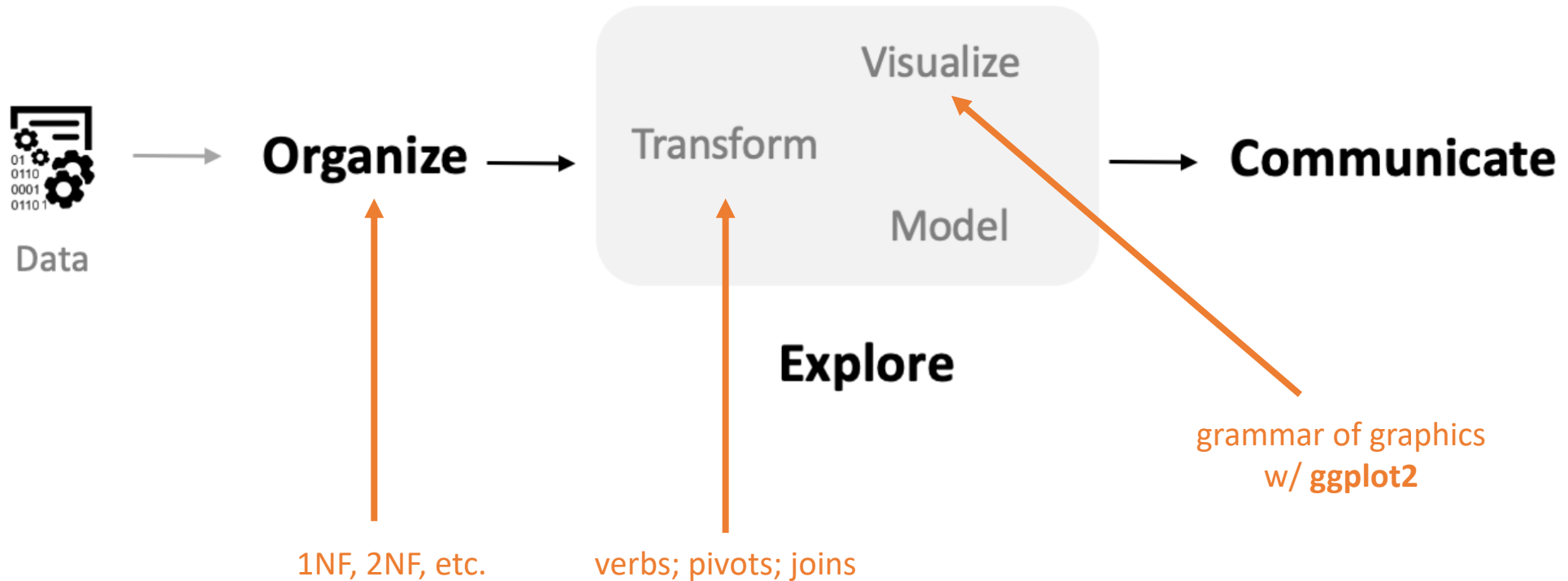
# Data Science Pipeline

A standard, highly abstract diagram showing the flow of information when doing data science work.



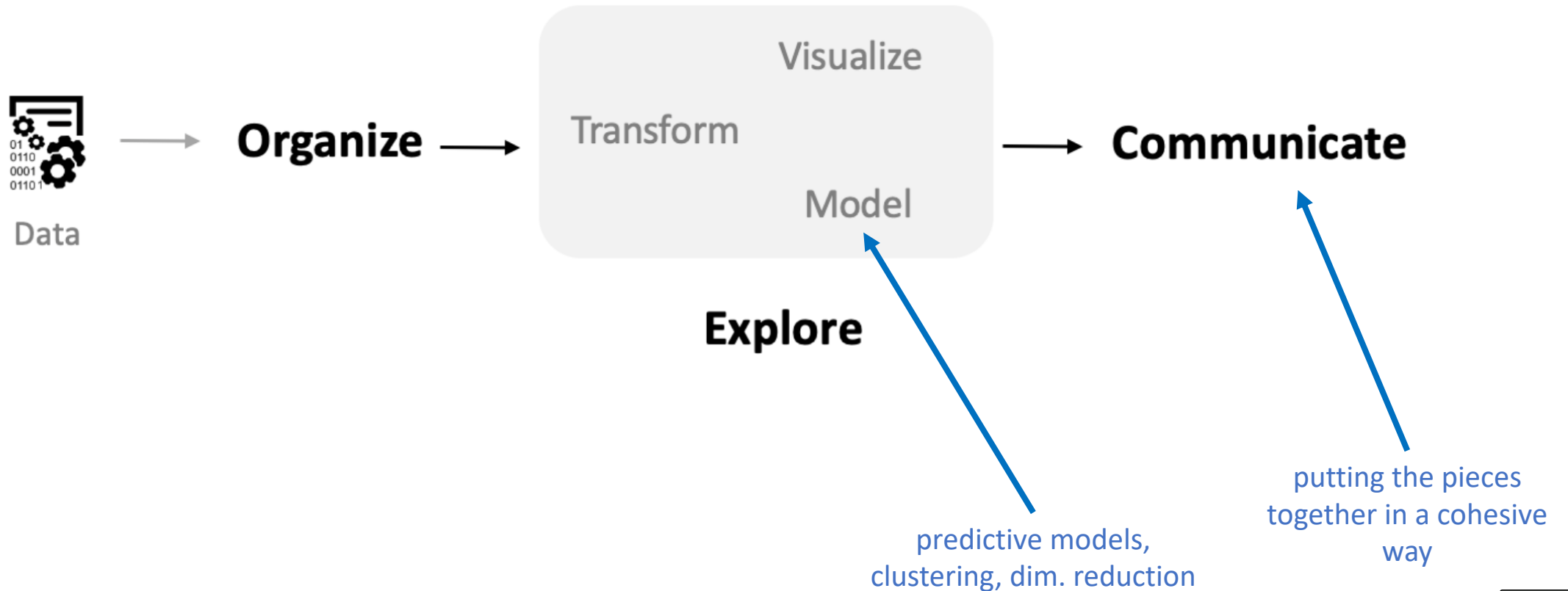
# DSST 289

In our Intro to Data Science course, we focused most heavily on the interior parts of the pipeline.



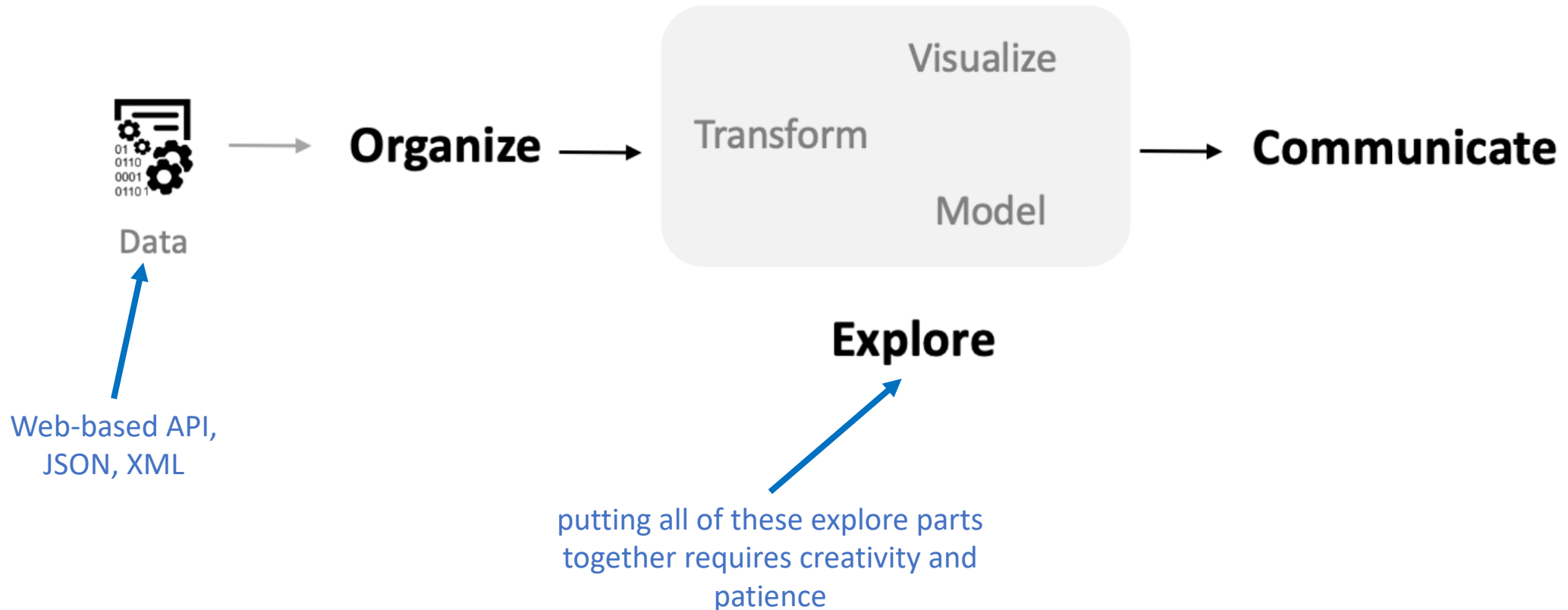
# DSST 389

For this class, we focused on the end of the pipeline while continuing to practice the interior methods.



# DSST 389

We also focus on the explore step as a whole and in the final project collecting data from an external API (a bit of a review for those in the Fall 2021 version of 289).



# Projects



# ML Concepts

## Core Concepts

- error rates
- train/validation split
- cross-validation
- confusion matrix
- overfitting

## Some Other Metrics

- false positive rate
- ROC curves / AUC
- Top-k error rate
- test/holdout error
- gain/lift

# Supervised Models

## Core Techniques

- linear regression
- logistic regression
- elastic net
- k-nearest neighbors
- decision trees
- gradient boosted trees

## Others You May See

- support vector machines
- additive models
- Bayesian models
- structural equation models
- kernel techniques



# Unsupervised Models

## Core Techniques

- principal components
- UMAP
- k-means
- hierarchical clustering

## Some Others You May See

- neural network embedding
- spectral clustering

*The choice of distance metric as a large effect on unsupervised models. We used Euclidean and TF-IDF. Many other techniques will modify the distance function but use a classical technique on the modified data*

# Text-Specific Techniques

## Core Techniques

- tokenization
- lemmatization
- POS tagging
- N-grams
- Topic Models
- KWiC
- G-scores

## Some Others You May See

- stochastic processes
- word2vec
- recurrent neural networks
- transformers (i.e., GPT-4)

# Data Creation

## Core Concepts

- APIs and HTTP
- Web Scraping
- XML & Xpath
- JSON & map functions
- iteration

## Some Others You May See

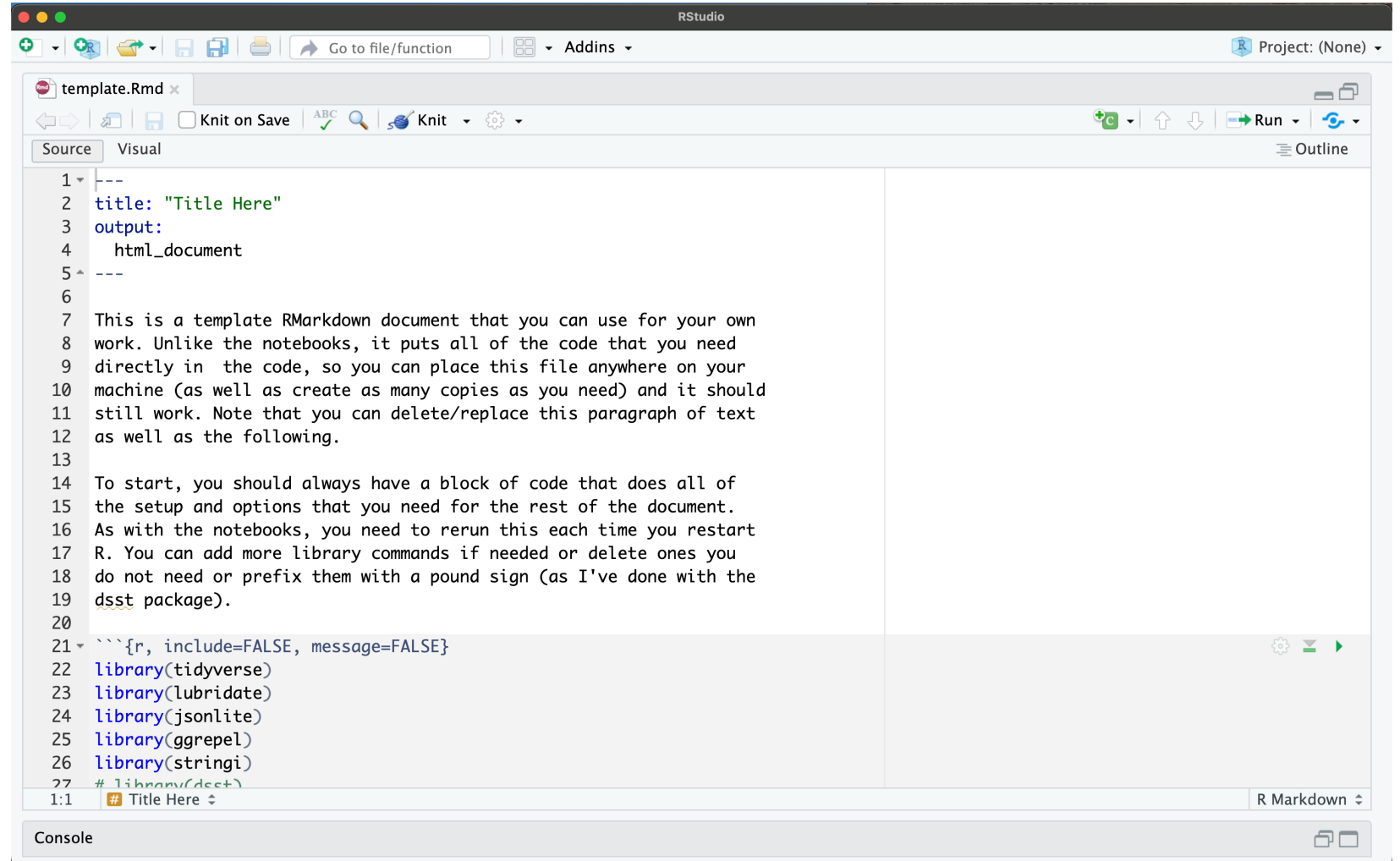
- regular expressions
- functional programming
- authentication
- cookies

*We only covered the basics of HTTP, XML, and JSON. You'll have enough to get started and can use other tutorials and documentation if you need to do something more complicated.*

# On Your Own

I have included a template file that you can get in our class notes. You can copy and use this file to run any code from class independently of the rest of the class setup.

The class notes will also remain online indefinitely in case you want to return to them.



```
1 ---
2 title: "Title Here"
3 output:
4   html_document
5 ---
6
7 This is a template RMarkdown document that you can use for your own
8 work. Unlike the notebooks, it puts all of the code that you need
9 directly in the code, so you can place this file anywhere on your
10 machine (as well as create as many copies as you need) and it should
11 still work. Note that you can delete/replace this paragraph of text
12 as well as the following.
13
14 To start, you should always have a block of code that does all of
15 the setup and options that you need for the rest of the document.
16 As with the notebooks, you need to rerun this each time you restart
17 R. You can add more library commands if needed or delete ones you
18 do not need or prefix them with a pound sign (as I've done with the
19 dsst package).
20
21 ```{r, include=FALSE, message=FALSE}
22 library(tidyverse)
23 library(lubridate)
24 library(jsonlite)
25 library(ggrepel)
26 library(stringi)
27 # library(dsst)
1:1 # Title Here
```

# DSST389: Statistical Learning

~~DSST389: Statistical Learning~~

**DSST389: Advanced Data Science**