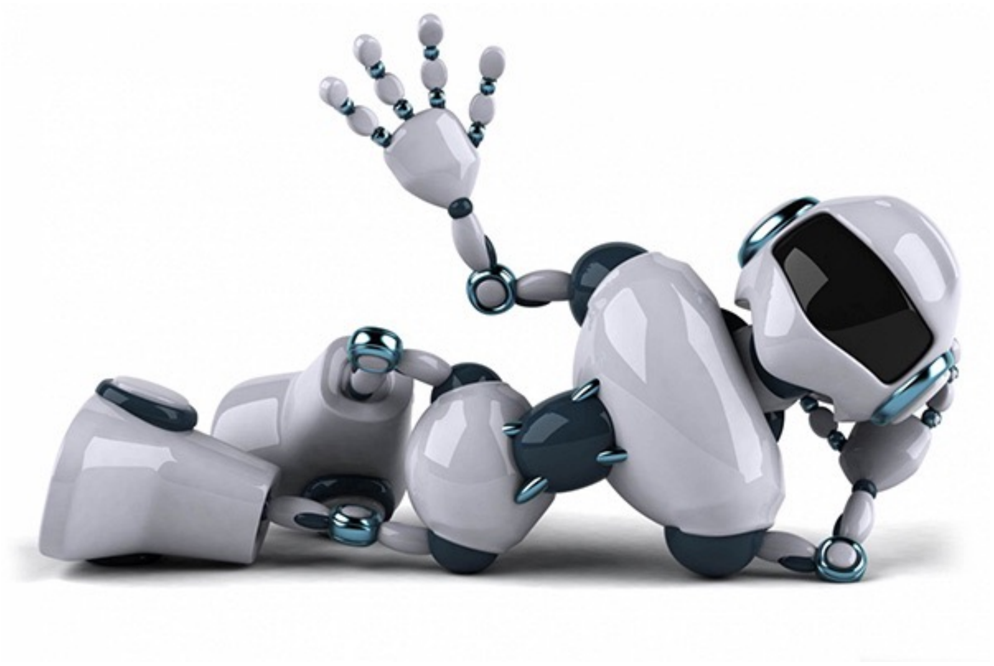
DSST389: Statistical Learning

Welcome!



T. ARNOLD

1. Syllabus

Course Website

As I mentioned in my email, these notes and all others for this semester will be posted on the course website:

https://statsmaths.github.io/dsst389-s23/

Course Expectations

To get right to the point, there are four main things I expect from you this semester:

- Regularly attend and participate in class.
- Record daily attendance (or reasons for absence) using the class form.
- Complete and present four class projects throughout semester.
- Complete end-of-semester self assessment of your work.

Grading

I try to keep grading simple. You will get a grade out of 95 points (based on the posted rubric) for each of the projects and another one for the self-assessment. The final grade is based on the average of these five scores.

Letter grades are assigned as follows: A (93–95), A- (90–92), B+ (87–89), B (83–86), B- (80–82), C+ (77–79), C (73–76), C- (70–72), and F (0–69).

Attendance

This is a course where it is very important to be present in class.

Please fill out the class form on the website at the **start of each class**. Note that there is a slight change from previous semesters. If absent, please explain why. I will follow up with anyone with a warning if there are any issues.

If you need to miss the day you are supposed to present a project, I expect you to (1) email me before class and (2) send me your slides.

Schedule and Workload

The course schedule is posted on the website. I will make every attempt to follow this schedule. I tried hard to avoid projects being due during typical busy periods.

The workload for this class is not particularly heavy but it is a bit inconsistent. Make sure you plan on carving out some time before the projects are due.

Getting Help

There is usually a lot of time during class and right after class to ask questions and get help with the course material.

I can answer quick questions by email. This is particularly helpful if you have a coding question.

Of course, I am also happy to set up a time to meet outside of class. I don’t have fixed office hours, but generally am free to meet on Mondays and Wednesdays after 1:30pm and before 5:30pm. Just send me an email (ideally the day beforehand) with some times that work for you.

Class Groups

This semester, I would like you all to organize yourselves into eight (or fewer) groups with between 2 and 4 students in each group. We will arrange the tables so that you are sitting with your group; you will work and/or share your results together during class.

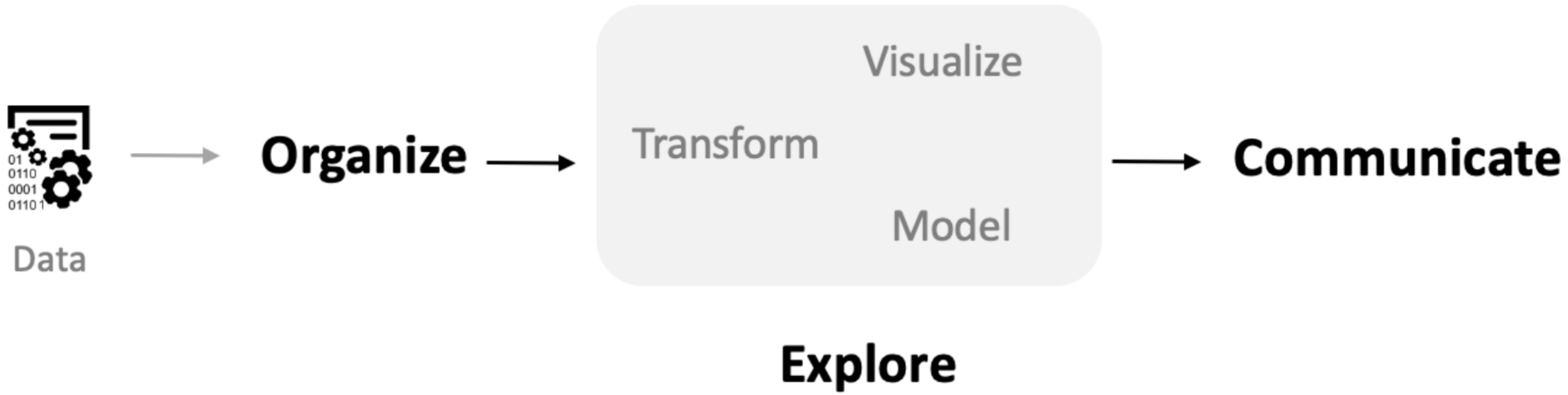
You have the **option** of submitting a joint project with your class group or any subset of your class group (maximum 3 students) for the first three projects. I recommend working in a pair if possible.

We will form these groups in a few moments.

2. Course Content

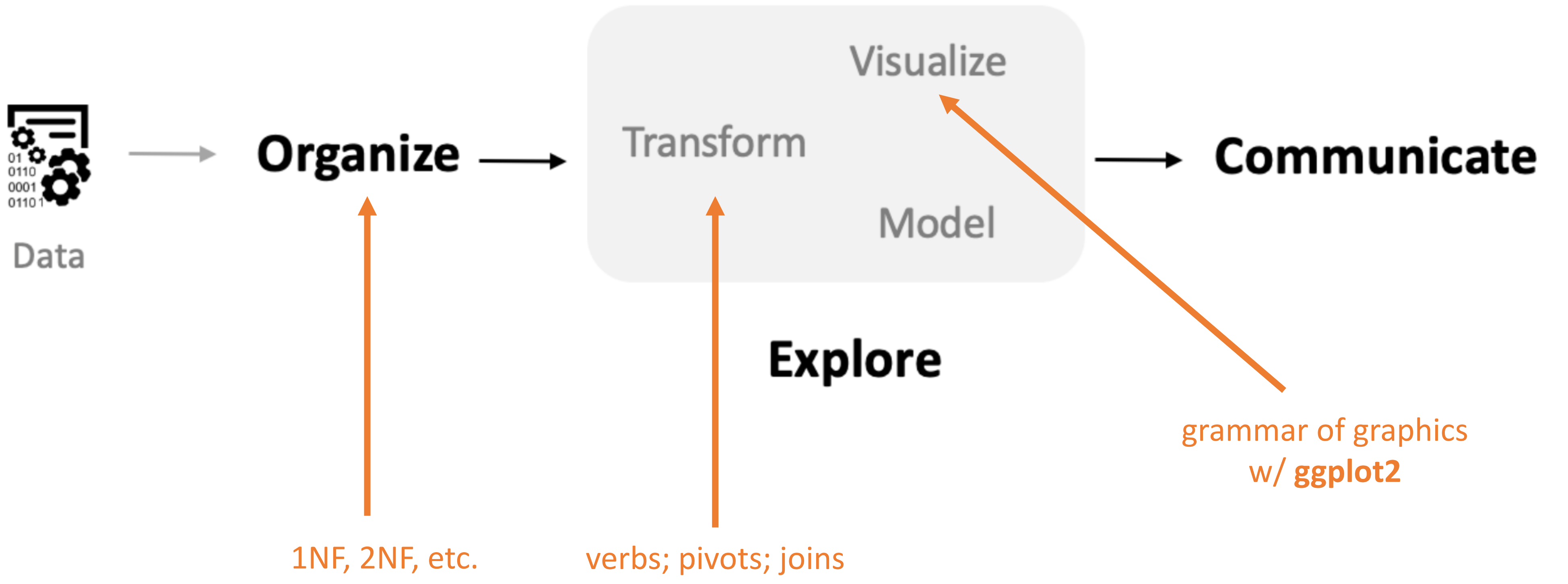
Data Science Pipeline

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



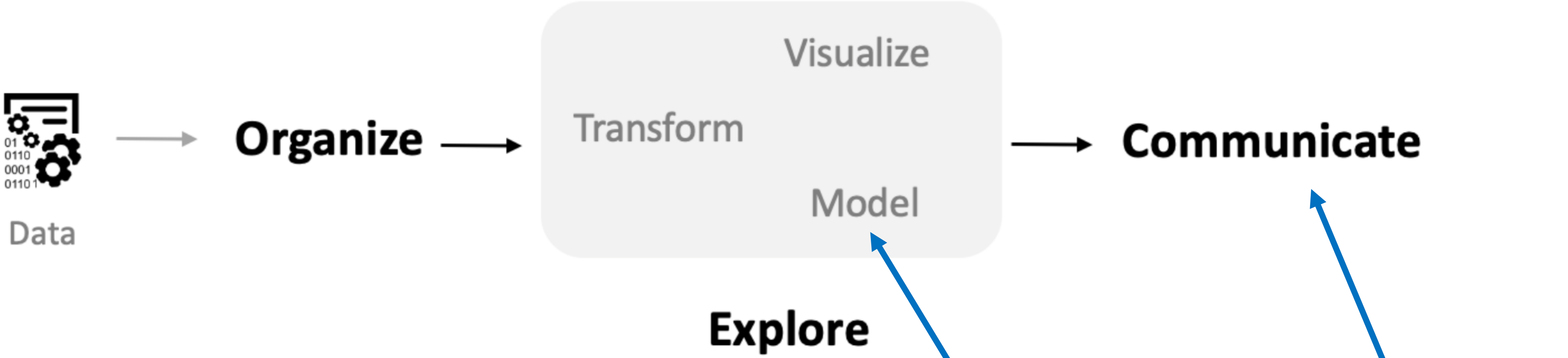
DSST 289

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



DSST 389

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

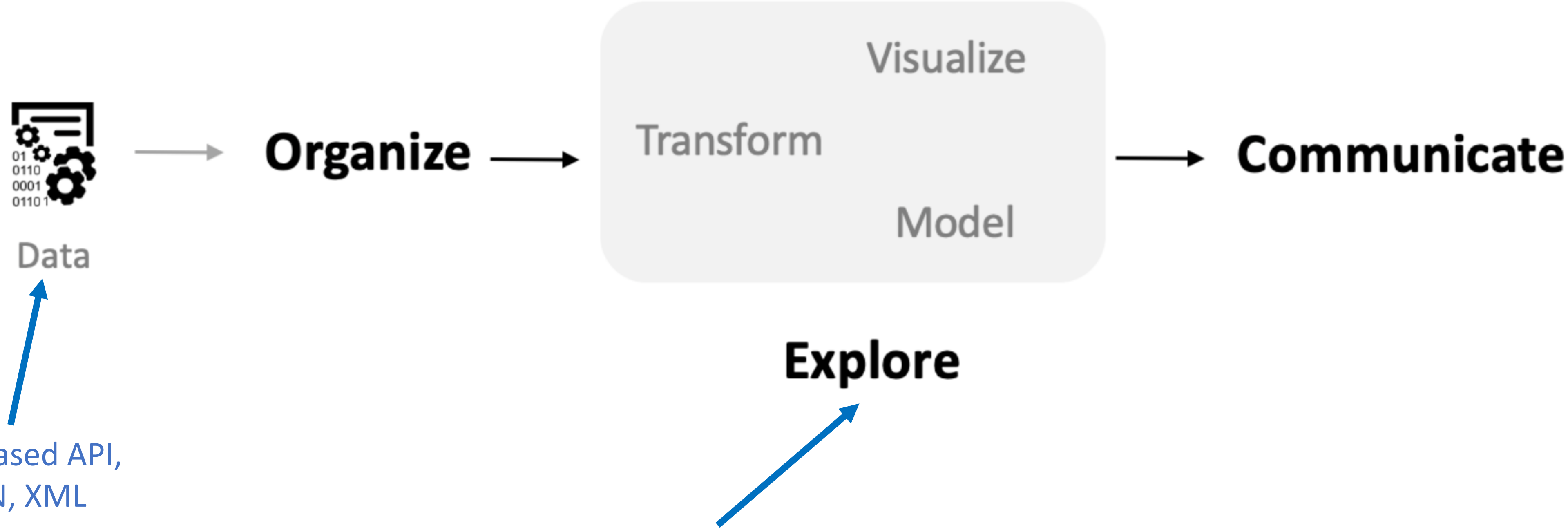


predictive models, clustering, dim. reduction

putting the pieces together in a cohesive way

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).



Web-based API,
JSON, XML

putting all of these explore parts
together requires creativity and
patience

What Kinds of Models?

Machine Learning vs *Statistical Learning*: different histories but the same thing

Machine Learning (ML) is a branch of artificial intelligence that uses data to create models. Methods mostly fall into three groups:

- **Supervised Learning**: detect patterns in order to make predictions about new data
- **Unsupervised Learning**: detect patterns in order to organize and structure data
- **Reinforcement Learning**: detect patterns in order to make complex decisions

These are not disjoint areas, though, many tasks require a mixture of methods.

Examples

Here are some examples of ML tasks. They are organized by the canonical type of machine learning that each is usually associated with, though the optimal approach may involve a mixture of methods.

Supervised Learning

- Predict the sale price of a house based on its size and location.
- Predict whether an email message should be put into a user's spam box.
- Find and identify all the faces found in a video feed.

Unsupervised Learning

- Cluster a collection of news paper articles by themes.
- Find and recommend similar products to users on a digital commerce website.
- Flag suspicious product reviews that should be manually investigated for fraud.

Reinforcement Learning

- Build an algorithm to play a game, such as checkers or chess, against a human.
- Determine a way to optimally schedule elevators in a large office building.
- Program a self-driving car.

What will you learn?

- understand the terminology of predictive and unsupervised ML methods
- how to apply a set of methods using the open-source R programming language
- how to use and understand a core set of general-purpose, interpretable ML methods
- how to use and understand several specific methods for working with textual data
- how to summarise and present the results of an exploratory analysis of data that integrates ML methods

What won't you (directly) learn?

- a laundry-list of dozens of ML methods
- theoretical justification / analysis of ML methods
- implementation details of ML methods
- deep learning models

Project Oriented Class

In order to embody the data science approach, this course is centered around four projects.

The primary goal will be learning how to use machine learning algorithms to tell a story about data. We will get a lot of practice doing this.

Our primary focus will be on text analysis. In addition to being generally interesting and one of my research areas, it is a great application domain for exploring the entire data science pipeline without any specialized domain knowledge.

Project Format

The projects take the form of a short presentation. Rather than a textual write-up, I want you to focus on producing clean, professional slides for the presentation.

Here are the planned topics:

- **IMDb movie reviews**: predicting how many stars a movie review gives
- **Amazon product reviews**: predict the author of a review
- **Yelp reviews**: predict author of the review and cluster the corpus authors
- **Wikipedia**: detect themes in a subset of Wikipedia articles

First Two Weeks

The content next two classes (remember that we don't have class on MLK day) will be a bit different. You'll be getting a crash course in the language of machine learning and we won't be doing any computing. Because of the dense nature of the material, I have prepared video notes in addition to the standard slides that you can watch, pause, and re-watch. In class we'll work on pen and paper handouts instead of R notebooks.

The material will be more mathematical than what I normally teach and may seem overwelming. Just do your best, watch the videos, and come ready to ask questions. From the third week onwards, 389 will feel a lot more similar (in fact, it's usually more fun) than what we did in 289.

3. Class Form + Groups + Setup