# Machine Learning in Practice

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

# What do you want/expect to learn from this class?

# Why does this class exist?

# What we want you to learn from this class

- How to responsibly and effectively solve real-world problems using ML
  - Understand the *entire* Machine Learning process (and get hands-on experience doing most of it)
  - Build (and use) reusable ML pipelines
  - Learn how to formulate ML problems, use, understand, evaluate, and communicate ML methods (that you have covered in earlier classes) in the context of a real problem

# How is this course different than typical ML classes you've taken before?

We'll assume everyone knows

- Methods/algorithms/models
- Assumptions behind them
- How to implement them

And focus on everything else that comes before the matrix and after the models are built (99% of the work done in a real-world project)

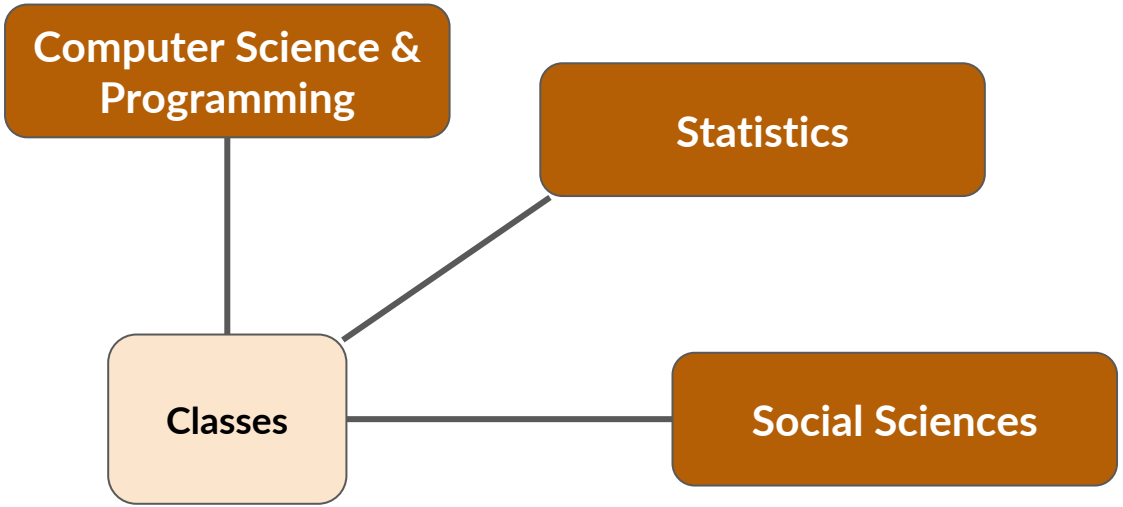
# Grading

- Pass/Fail
- We want the focus to be on learning and not on the grade.
- Levels of learning
  - Exposed to information covered in this course
  - Applying the information (correctly) covered to the class project
  - Generalizing to the next ML problem you tackle

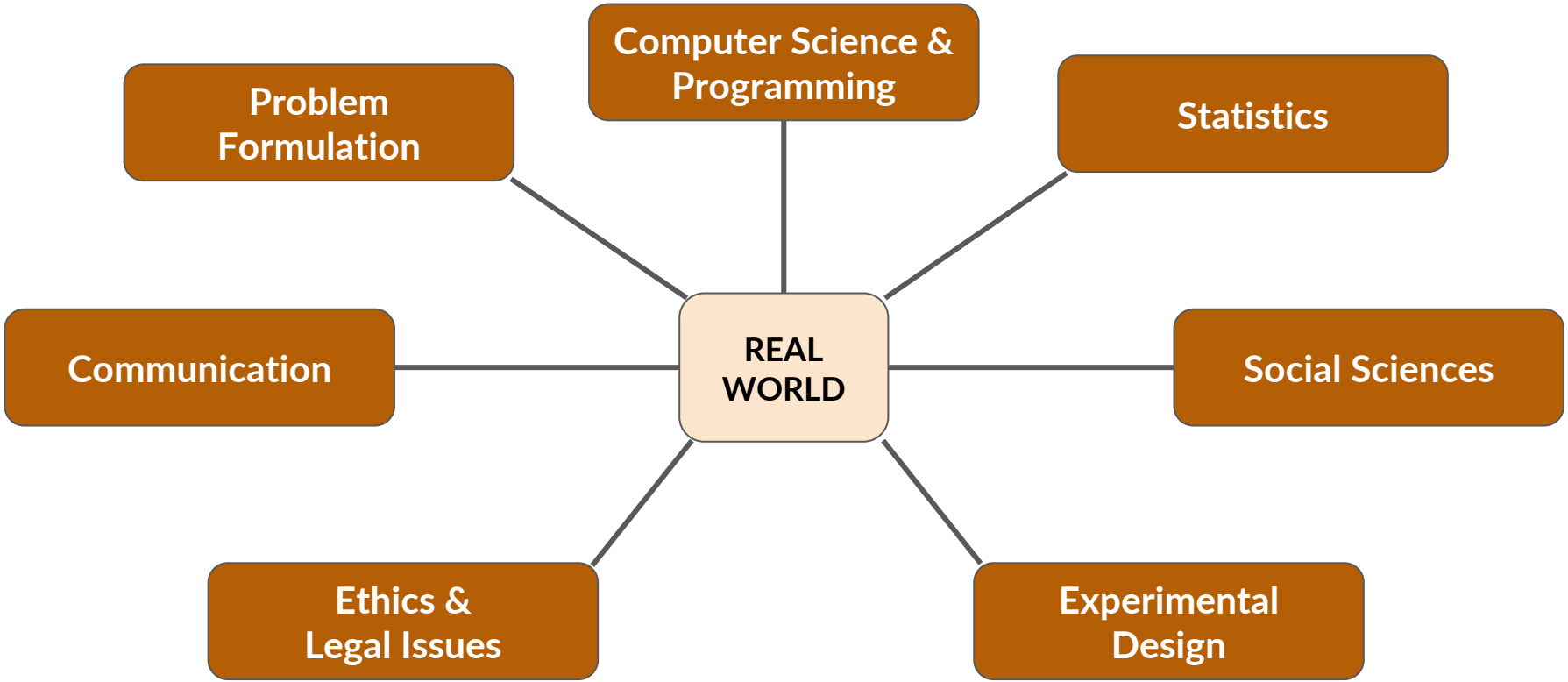
# Pre-requisites

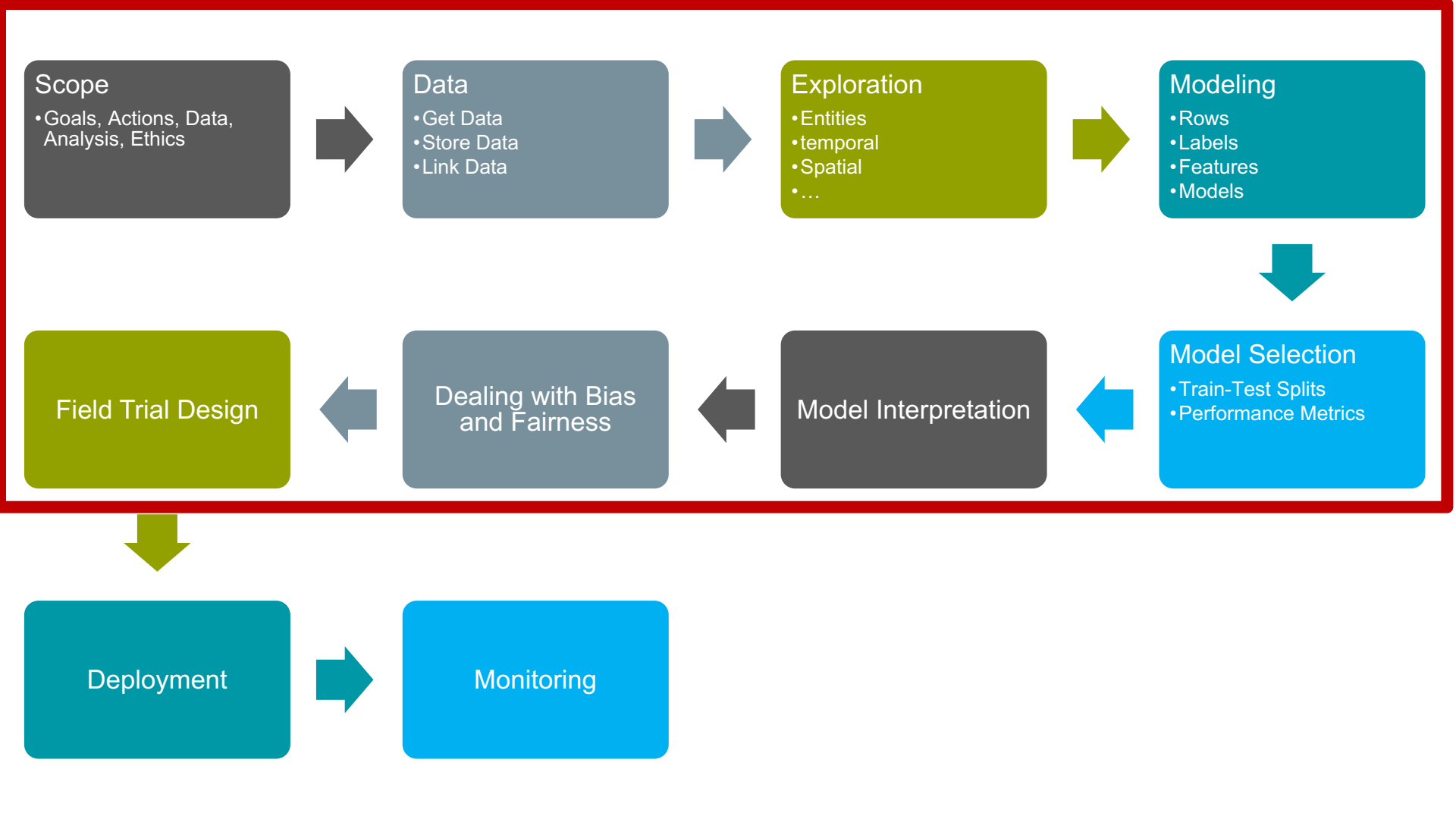
- Machine Learning (methods and overall process)
- Python (pandas, sklearn, tensorflow, matplotlib)
- Ideally: experience with SQL, command line (bash), git(hub), working on remote servers

# Skills needed to solve real-world problems (with ML)



# Skills needed to solve real-world problems (with ML)




Scope
•Goals, Actions, Data, Analysis, Ethics
Data
•Get Data
•Store Data
•Link Data
Exploration
•Entities
•temporal
•Spatial
•…
Modeling
•Rows
•Labels
•Features
•Models
Model Selection
•Train-Test Splits
•Performance Metrics
Model Interpretation
Dealing with Bias and Fairness
Field Trial Design
Deployment
Monitoring

# Structure of the class

- Module 1: End-to-end ML Pipeline
  - Formulation, Modeling Setup, Features, Models, Model Selection
- Module 2: Model Interpretability
- Module 3: Fairness

# Class Schedule

| Week | Dates | Topic | Required Readings | Assignments |
|---|---|---|---|---|
| 1 | Tu: Aug 31 | Class Intro and Overview | | |
| 1 | Th: Sep 2 | ML Project Scoping | ML Project Scoping Guide | Project Team Selection |
| 2 | Tu: Sep 7 | Getting, Storing, and Linking Data | Optional readings on github | |
| 2 | Th: Sep 9 | Analytical Formulation / Baselines | List on github | |
| 3 | Tu: Sep 14 | Model Selection Methodology | | Project Assignment 1: Formulation and Baseline (due Monday) |
| 3 | Th: Sep 16 | Performance Metrics | | |
| 4 | Tu: Sep 21 | Feature Engineering and Imputation | | Project Assignment 2:<br>Validation set up<br>Initial pipeline with train and validation set(s) and baseline implemented (due Monday) |
| 4 | Th: Sep 23 | Hands-on Session for ML Pipeline review | | |
| 5 | Tu: Sep 28 | Models/hyperparameters in practice | | Project Assignment 3:<br>list of features and some subset implemented (due Monday) |
| 5 | Th: Sep 30 | Temporal Model Selection | | |
| 6 | Tu: Oct 5 | Module 1 Review: Applied ML - End to End Pipelines | | Project Assignment 4:<br>modeling results (due Monday) |
| 6 | Th: Oct 7 | Mid-term week - no class | | Mid-term exam due Friday |
| 7 | Tu: Oct 12 | Interpretability: Intro and Overview, taxonomy | | |
| 7 | Th: Oct 14 | No Class - Mid-semester break | | |

# Class Schedule

| Week | Dates | Topic | Required Readings | Assignments |
|---|---|---|---|---|
| 8 | Tu: Oct 19 | Understanding the Models | | |
| 8 | Th: Oct 21 | Interpretability Methods: Inherently Interpretable (GA2Ms, RiskSLIM, etc.) | | |
| 9 | Tu: Oct 26 | Interpretability Methods:: Post-Hoc Local/Feature-based (LIME, SHAP, MAPLE) | | |
| 9 | Th: Oct 28 | Interpretability Methods: Other methods (counterfactual, example-based, etc.) | | |
| 10 | Tu: Nov 2 | Module 2 Review: ML Interpretability | | |
| 10 | Th: Nov 4 | ML Ethics Issues Overview | | Interpretability Writeup Due on Friday |
| 11 | Tu: Nov 9 | Fairness in ML Overview | | |
| 11 | Th: Nov 11 | Fairness Methods: Pre-processing (removing sensitive attribute, sampling) | | |
| 12 | Tu: Nov 16 | Fairness Methods: In-processing (Zafar, Celis, fairlearn, etc.) | | |
| 12 | Th:Nov 18 | Post-Processing: Hardt, LA, etc | | |
| 13 | Tu: Nov 23 | Module 3 Review: ML Fairness | | |
| 13 | Th: Thanksgiving | Thanksgiving holiday | | |
| 14 | Tu: Nov 30 | Field Trials and Causality | | Bias Writeup Due |
| 14 | Th: Dec 2 | Wrap-Up | | |
| 15 | Tu: Dec 7 | No Class - Finals Week | | |
| 15 | Th: Dec 9 | No Class - Finals Week | | Final Research Writeup Due |

# Logistics

- Attendance (is not optional)
- Platforms:
  - Latest content will be on **github**
  - **Canvas** (for assignment submissions)
  - **Slack (and email)** for communications and project and teamwork
- Wednesday sessions
- Office hours
- TAs: to help manage and help access AWS infrastructure

# Project Teams

- Make sure to fill out the survey: We need your github username and  public SSH key

- You should create 5-person teams by the end of this week (and fill out the spreadsheet to let us know)

# Project

## Support a classroom.
## Build a future.

DONORS CHOOSE

Teachers and students need your support more than ever. Get crayons, books, cleaning supplies, technology, and more to help students get back to learning.

See classroom projects

Our efficiency and transparency have earned us the highest rating on Charity Navigator.

# Tech Setup Options

- Get set up on github for your project
- Compute Infrastructure
  - Use your own laptop/machine for running things locally
  - Or use AWS resources we have set up to run larger jobs
- Data Infrastructure
  - Use CSVs on your own laptop/machine
  - **Or use Postgres database we have set up with data loaded**
- If you use AWS, make sure you have the following things set up:
  - ssh (to connect to the server) **server.mlinpractice.dssg.io**
  - dbeaver and psql (to connect to the database) **server.mlinpractice.dssg.io**

- Nice to have: get familiar with
  - Postgresql (to analyze and query data)
  - *nix command line
  - Remote server workflow

# Prep for next class

- Reading
  - Project Scoping guide
- Assignments
  - Survey (if you haven’t already): link from email
  - Project Team Selections: link in canvas
- No Wednesday Session This Week