CIS 520: Machine Learning

Project Introduction / Guidance

Barry Plunkett

Department of Computer & Information Science
University of Pennsylvania

Spring 2019

Outline

- Introduce Project
  - Goals
  - High-level overview
  - Deliverables, deadlines, and evaluation
- Guidelines
  - Picking a dataset
  - Optimal timeline
  - Spotlight slides and report examples

Project Goals

- Design and implement a machine learning approach to solving a real-life problem
- Apply theory from class to justify your modeling decisions
- Possibly draw on theory to modify existing algorithms when necessary

Overview

1. Choose a problem and a dataset
2. Conduct a literature review of solutions
3. Frame the project as a machine learning problem (supervised, unsupervised)
4. Clean and pre-process data into usable format
5. Choose several solution algorithms to implement
6. Write project report and give spotlight presentation

Deliverables and Deadlines

Deliverable	Due Date
Proposal	March 12
Milestone Meeting	April 5
Presentation	April 26
Report	April 30

Proposal

Roadmap outlining problem you plan to solve and how you will solve it

Deliverables and Deadlines

Deliverable	Due Date
Proposal	March 12
Milestone Meeting	April 5
Presentation	April 26
Report	April 30

Milestone Meeting

Meeting with project mentor to discuss progress and challenges

Deliverables and Deadlines

Deliverable	Due Date
Proposal	March 12
Milestone Meeting	April 5
Presentation	April 26
Report	April 30

Presentation

Brief presentation giving high-level overview of problem and results

Deliverables and Deadlines

Deliverable	Due Date
Proposal	March 12
Milestone Meeting	April 5
Presentation	April 26
Report	April 30

Report

Thorough 5-page analysis of methodology and findings

Evaluation Criteria

- Technical Quality
- Novelty
- Clarity of Presentation
- Significance

Guidelines – Choosing a Dataset

Characteristics

- Simple to Wrangle
- Moderate Size
- Few Missing Values
- Answers Question

Simple to Wrangle

- Downloadable as a .csv, .json, or text file
- Accessible via a free, well-documented API

Tip: No bonus points for creative scraping or hand-labeling

TIP: If you're building a dataset using an API, finish building shortly after submitting proposal

Guidelines – Choosing a Dataset

Characteristics

Simple to Wrangle

Moderate Size

Few Missing Values

Answers Question

Moderate Size

- Large enough to learn complex models
- Small enough to load into memory
  - 8 GBs RAM : < 4 GBs data
  - 12 GBs RAM : < 8 GBs data
  - 16 GBs RAM: < 12 GBs data

Tip: Don’t expect BigLab will empower you

TIP: Only plan to rely on cloud computing platforms (AWS, Google Cloud, Azure) if you have experience

Guidelines – Choosing a Dataset

Characteristics

- Simple to Wrangle
- Moderate Size
- Few Missing Values
- Answers Question

Few Missing Values

- Most datasets have missing and misentered values
- Too many can make learning more challenging
- Many imputation methods exist

Tip: Determine how flawed your data is **before** anything else

TIP: Avoid datasets with **> 15%** flawed/missing entries

Guidelines – Choosing a Dataset

Characteristics

- Simple to Wrangle
- Moderate Size
- Few Missing Values
- Answers Question

Answers Question You Pose

- For many questions, finding a dataset with meaningful features for learning can be challenging
- Bear this in mind when choosing a problem

Tip: Start with a "good" dataset, then think of questions

TIP: Formulate hypotheses for why features should cause or predict the outcome of interest before choosing a dataset

Guidelines – Choosing a Dataset

Characteristics
Simple to Wrangle
Moderate Size
Few Missing Values
Answers Question

Caveats

- These are **guidelines, not requirements**
- Dataset may not satisfy these requirements
- If your dataset deviates, be aware of consequences
- Reach for the stars!

Guidelines – Timeline

Timeframe
Now – 3/12
3/12 – 4/5
4/5 – 4/12
4/12 – 4/19
4/19 – 4/26
4/26 – 4/30

Before Proposal

- Select problem and “good” dataset
- Load dataset into memory
- Compute summary statistics, measure quality
- Complete thorough literature review
- Set ambitious target for milestone meeting

TIP: Ask each group member to read at least two related papers and several blog posts at this stage

Guidelines – Timeline

Timeframe
Now – 3/12
3/12 – 4/5
4/5 – 4/12
4/12 – 4/19
4/19 – 4/26
4/26 – 4/30

Before Milestone Meeting

- Complete all data wrangling
- Clean data & impute missing values
- Feature engineering
- Split into train/test/dev sets
- Transform data into format expected by models
- Implement minimum viable solution

TIP: At this stage re-watch lectures and re-read notes related to chosen algorithms. Meet to resolve confusion.

Guidelines – Timeline

Timeframe
Now – 3/12
3/12 – 4/5
4/5 – 4/12
4/12 – 4/19
4/19 – 4/26
4/26 – 4/30

Week After Milestone

- Fully implement selected algorithms
- Train baseline versions of each model using arbitrary hyperparameters
- For each hyperparameter of each model, choose set of values to experiment with later

TIP: Measure training time of each model to gauge what types of hyperparameter searches will be feasible later

Guidelines – Timeline

Timeframe
Now – 3/12
3/12 – 4/5
4/5 – 4/12
4/12 – 4/19
4/19 – 4/26
4/26 – 4/30

Two Weeks After Milestone

- Train models for all hyperparameter settings
- Record dev set / cross validation performance
- Train models with optimal settings and record test set performance
- Generate plots of training / dev error

TIP: If possible, avoid training schemes that require multiple days or many hours and/or record results intermittently.

Guidelines – Timeline

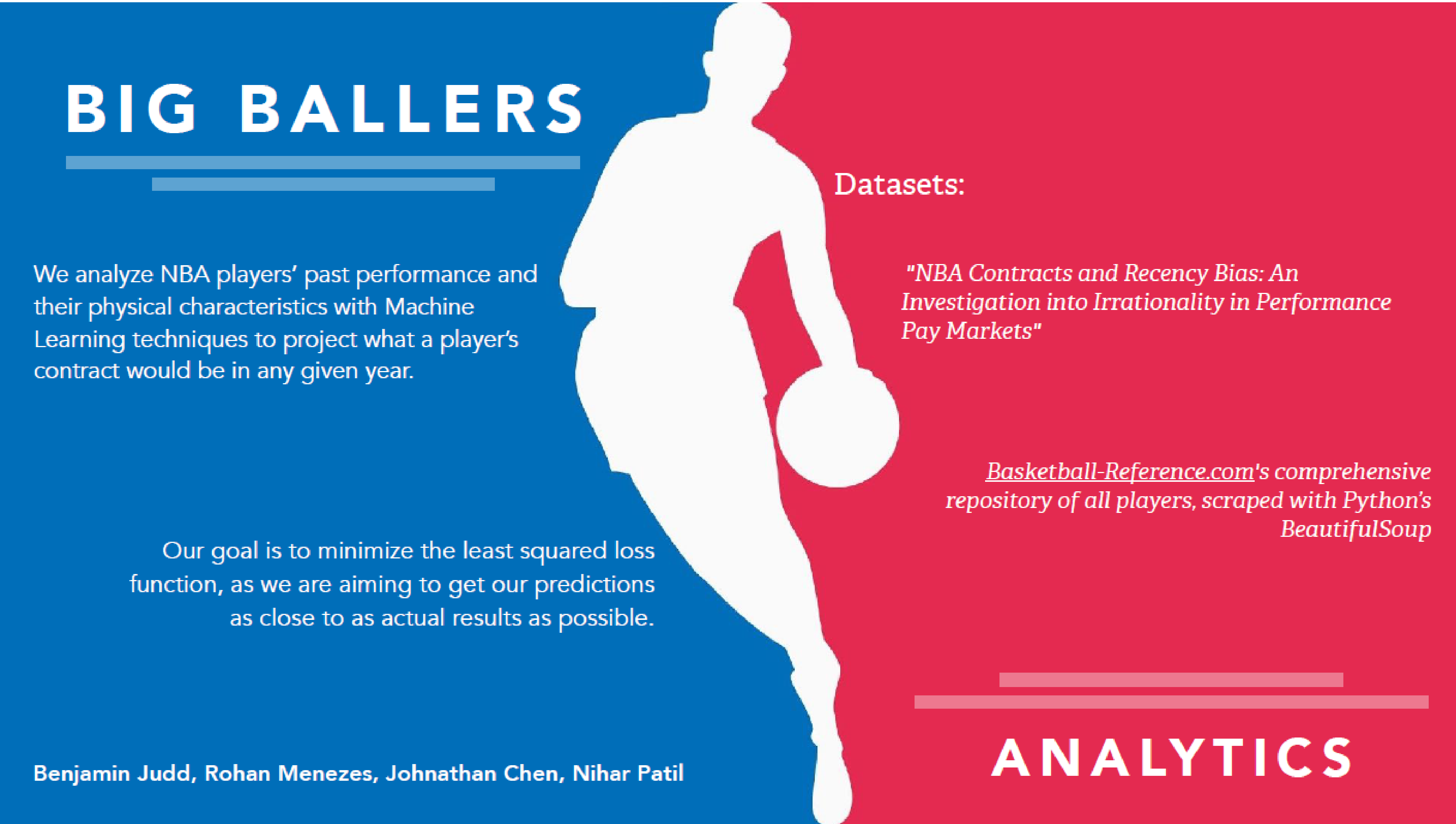
Timeframe
Now – 3/12
3/12 – 4/5
4/5 – 4/12
4/12 – 4/19
4/19 – 4/26
4/26 – 4/30

Week Before Presentation Due

- Finalize model selection and test set results
- Write first draft of project report
- Review notes from literature review
- Meet with group members to discuss conclusions
- Prepare spotlight presentation slides

TIP: Visually appealing spotlight slides help your project standout during grading

Spotlight Slides Example



BIG BALLERS

We analyze NBA players' past performance and their physical characteristics with Machine Learning techniques to project what a player's contract would be in any given year.

Our goal is to minimize the least squared loss function, as we are aiming to get our predictions as close to as actual results as possible.

Benjamin Judd, Rohan Menezes, Johnathan Chen, Nihar Patil

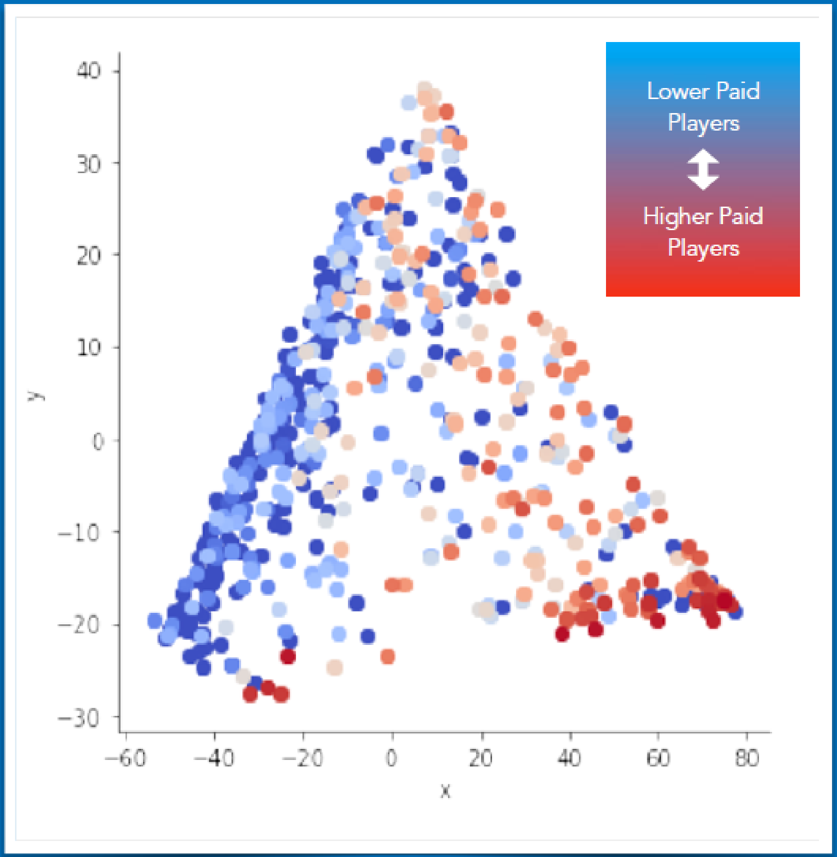
Datasets:

"NBA Contracts and Recency Bias: An Investigation into Irrationality in Performance Pay Markets"

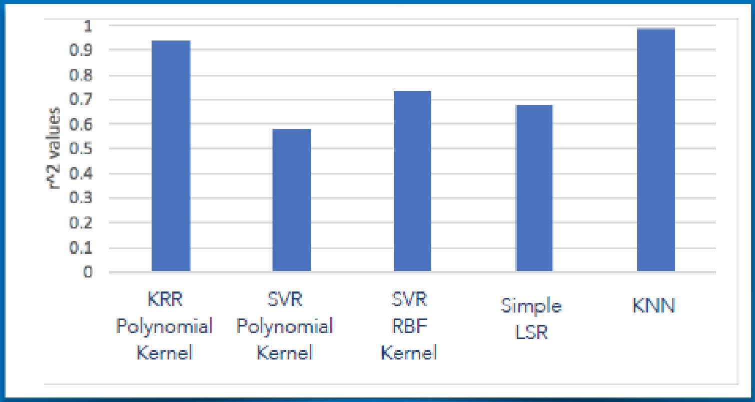
Basketball-Reference.com's comprehensive repository of all players, scraped with Python's BeautifulSoup

ANALYTICS

Spotlight Slides Example



Overall Results



- 90-10 train-test split for our models
- Used cross validation to select the degrees for polynomial kernels and the number of neighbors for NN
- Computed r^2 values by averaging over several splits

Guidelines – Timeline

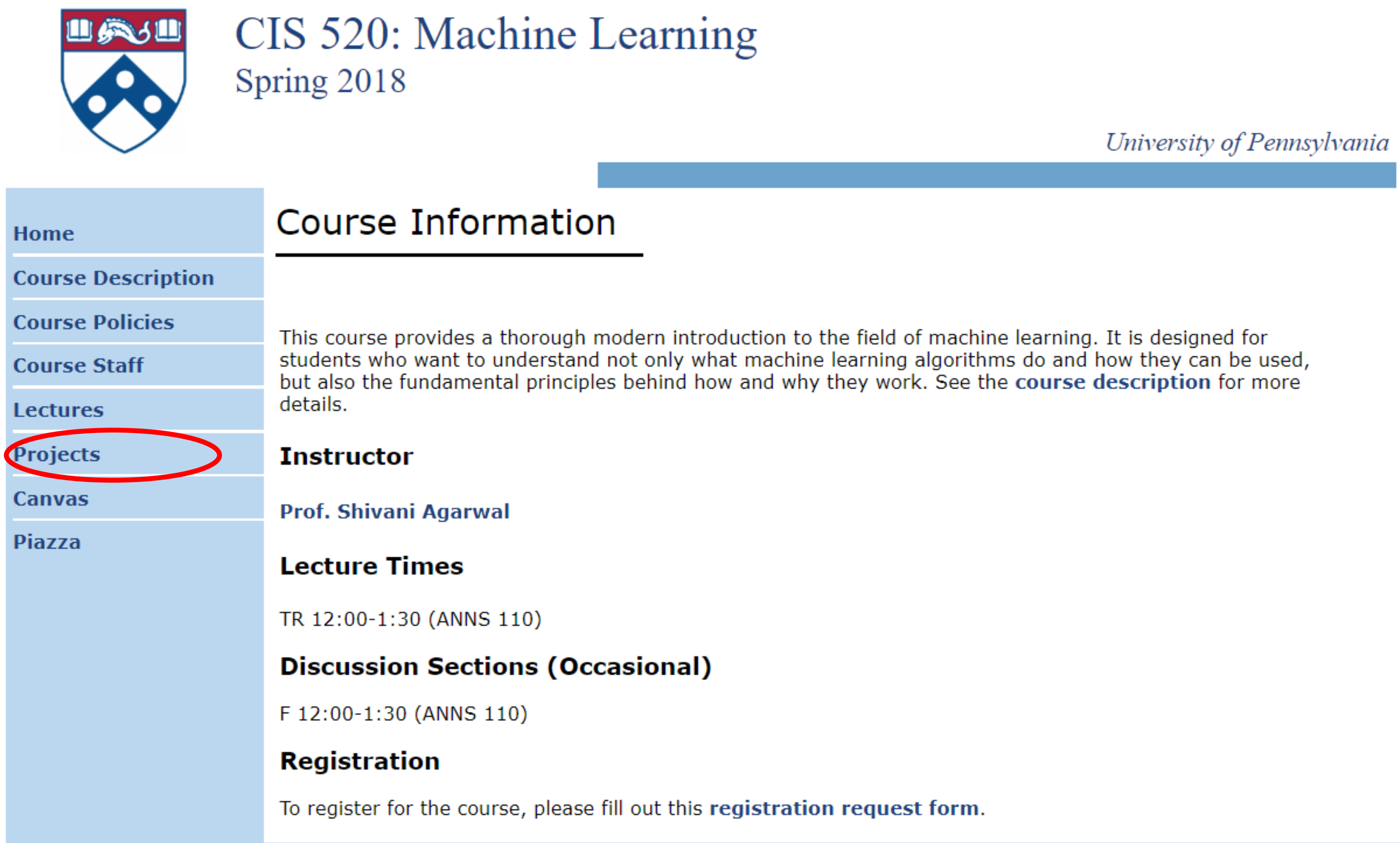
Timeframe
Now – 3/12
3/12 – 4/5
4/5 – 4/12
4/12 – 4/19
4/19 – 4/26
4/26 – 4/30

Final Days Before Report Due

- Rehearse spotlight presentation
- Re-write and proofread report
- Precisely define variables for all equations
- Prepare table of test set results
- Break up text with illustrations, visualizations, error plots, etc.

TIP: Have all sections of the report proofread by at least one member other than the section author

Report Examples



CIS 520: Machine Learning

Spring 2018

University of Pennsylvania

Course Information

This course provides a thorough modern introduction to the field of machine learning. It is designed for students who want to understand not only what machine learning algorithms do and how they can be used, but also the fundamental principles behind how and why they work. See the **course description** for more details.

Instructor

Prof. Shivani Agarwal

Lecture Times

TR 12:00-1:30 (ANNS 110)

Discussion Sections (Occasional)

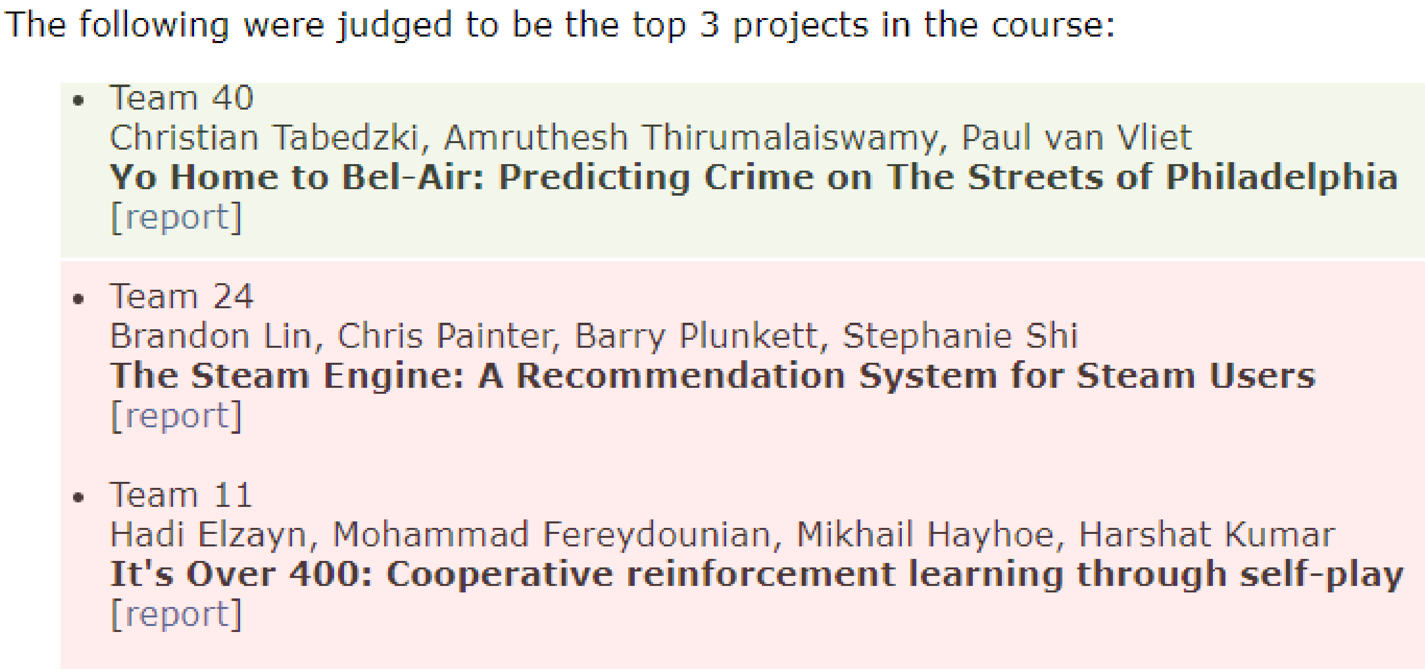
F 12:00-1:30 (ANNS 110)

Registration

To register for the course, please fill out this **registration request form**.

www.Shivani-Agarwal.org/Teaching/CIS-520/Spring-2018/

Report Examples



The following were judged to be the top 3 projects in the course:

- Team 40
  Christian Tabedzki, Amruthesh Thirumalaiswamy, Paul van Vliet
  Yo Home to Bel-Air: Predicting Crime on The Streets of Philadelphia
  [report]
- Team 24
  Brandon Lin, Chris Painter, Barry Plunkett, Stephanie Shi
  The Steam Engine: A Recommendation System for Steam Users
  [report]
- Team 11
  Hadi Elzayn, Mohammad Fereydounian, Mikhail Hayhoe, Harshat Kumar
  It's Over 400: Cooperative reinforcement learning through self-play
  [report]

Supervised Learning

Collaborative Filtering

Reinforcement Learning