DATA SCIENCE PRACTICUM

SPRING 2018 WRAP-UP LECTURE
OUTLINE

1. Some fun stats
2. Course feedback
3. Wrap-up
COMMITS

- 2,605 commits
- 7,347,331 additions, 2,906,254 deletions
- 4,441,077 new lines of code
NORMALIZED PROJECT PERFORMANCE

Takeaway:
Good AutoLab performance != High marks
TAKEAWAYS

• **Reuseability is critical** in academia and industry frameworks
  – Documentation
  – READMEs
  – Comments
  – Development process (Issues, Pull Requests, Gitter/Slack/Listservs)

• The real world operates in **teams**

• Reusability + teamwork = **effective communication is absolutely essential**
  – **More important** than Kaggle leaderboard position (what’s the point if only you can understand it)
  – **More important** than raw coding talent (raw coding talent is dime/dozen)
AUTOLAB SCORE PROGRESSION

Takeaway: Project difficulty increased / time availability decreased
INDIVIDUAL GRADE PROGRESSION

Takeaway: Improvement over time!
FEEDBACK

• Activity time!
• Three categories of feedback:
  1. **STOP**: What would you *remove* or *eliminate* from the course?
  2. **START**: What would you *add* to the course?
  3. **CONTINUE**: What would you *keep* or *retain* that is already present in the course?

• Individually, write down one item for each category (5 minutes)
• Get in groups and develop a consensus list (10 minutes)
FEEDBACK

MY THOUGHTS

• STOP
  – 4 projects + Final project is too much; cut down 1 course project, extend the others, leave more room for Final project

• START
  – New category of projects like unsupervised clustering, matrix factorization, reinforcement learning

• CONTINUE
  – Project 0 (maybe extend in length a tad)
• Teams aren’t going anywhere
• Lectures will be tweaked for additional background knowledge
  – Also, 3360 Data Science I is proposed to become a 4000/6000 course to allow graduate enrollment
  – When that happens, DS I will be a required prerequisite to 8360 (no more “toughing it out”)
• Switch from Spark to dask
• Submit programs to AutoLab, rather than just predictions

• EXTRA CREDIT: By midnight, April 27, propose a new project idea. Needs:
  – A clear, unambiguous ground-truth (or evaluation metric) to put in AutoLab
  – An openly available dataset (or one that can be acquired, e.g., CodeNeuro or cilia)
  – Can achieve a reasonable solution in 2-3 weeks
  – Up to 5 points on your final grade
OTHER THOUGHTS

- Course evaluations!

http://eval.franklin.uga.edu/
FINAL PRESENTATIONS

- **Wednesday, April 18**
  - Parya, Omid, Raunak
  - Jeremy, Ailing

- **Thursday, April 19**
  - Hiten, Ankit
  - Ankita, Vibodh, Vyom
  - Nihal, Vamsi, Vinay Bingi

- **Tuesday, April 24**
  - Chris, Zach
  - Jin
  - Prajay, Nick, Layton

- **Wednesday, April 25**
  - Weiwen, I-Huei
  - Rajeswari, Maulik

**Friday, April 27, 11:59pm:**
ALL PROJECT MATERIALS DUE
Final Notes

- Raw scores aren’t everything
- ...but they’re a reasonable indicator
- Balance exploration (examining the data, testing multiple approaches) with exploitation (designing, testing, and documenting a complete pipeline)
  - Corollary: start early!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

Questions?
THANK YOU!

THANK YOU

STAY CLASSY