Introduction to Statistics and Machine Learning in Astronomy

ASTR 324

mjuric@astro.washington.edu

This Week

- Why statistics and ML in astronomy?
- Administrivia
- Getting set up (Slack, JupyterHub, GitHub,…)

About Me: Prof. Juric

Artist's impression: 👇

- Mario Juric (mar-ee-oh you-rich)
  - Astronomy Prof & eScience Institute Fellow
  - Office: PAC 320
- What I do:
  - Rubin Observatory Legacy Survey of Space and Time (LSST)
  - Astronomical algorithms and software research
  - Science derived from large surveys: Galactic structure, properties of the Solar System

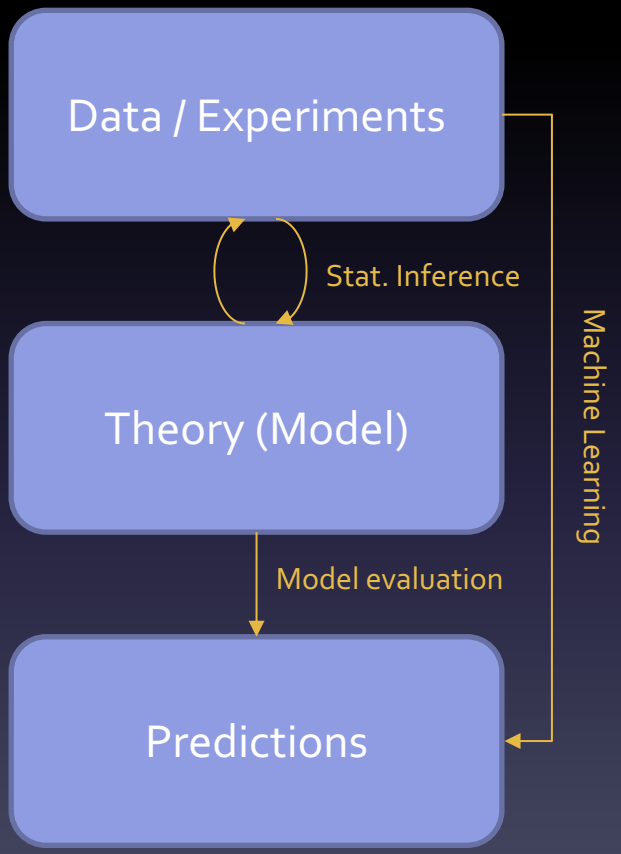
Photographic representation: 👈

Why Statistics and ML?

- Our goal as scientists is to *understand* the laws governing the world around us (theories) based on *observations and experiments* (data), and make *predictions* about yet unseen observations and experiments.

- Statistics (statistical inference) gives us the mathematics to correctly interpret observations and their impact on theories.

- Machine learning allows us to sometime (in part) skip the "understanding" part, jumping straight from data to predictions. Interpreting ML models (understanding the *"why?"*) is one of the most interesting areas of ML today.



A new topic every week

1. Getting started with technology
2. Introduction to probability and statistics I
3. Introduction to statistics II
4. Maximum likelihood and applications in astronomy
5. Bayesian inference and model selection
6. Introduction to MCMC and model parameter estimation
7. Dimensionality reduction
8. Time series analysis
9. Machine learning I
10. Machine learning II

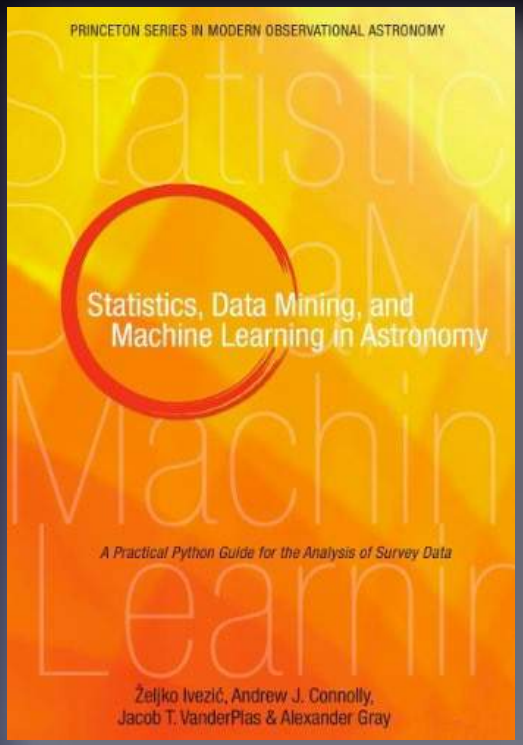
https://github.com/uw-astr-324/astr-324-s22/blob/master/syllabus/syllabus.pdf

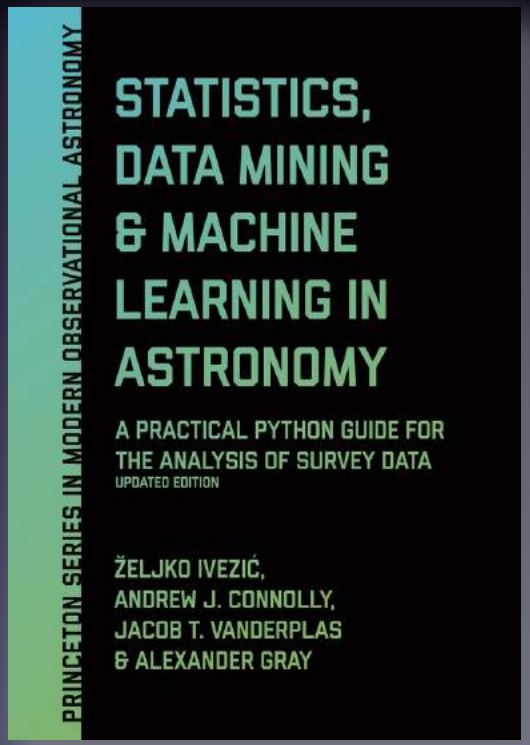
Learning Goals

- At the end of this course, you should know:
  - How to think probabilistically, and correctly interpret probability
  - How to correctly summarize measurements
  - How to estimate model parameters given observations, and when to reject poor models. Understand the theory behind why this works.
  - How to think of probabilities of parameter values, and how to derive those using Markov Chain Monte Carlo techniques.
  - How to measure and interpret time series
  - What is machine learning, and how to apply it.
  - What machine learning is not, and when not to apply it.

Textbooks & Reference Material

- We will be using the "Statistics, Data Mining, and Machine Learning in Astronomy" textbook by Ivezic, Connolly, VanderPlas and Gray





+ many (many) online writeups / notebooks / blog posts that I will point you to over the next few weeks.

Class Meetings

- When: TTh, 10am-11:20am
- Were & How:
  - Lectures delivered via YouTube: https://dirac.us/videos324
    - Generally Jupyter notebooks; best to follow along as you watch.
    - Supplement by readings from the textbook.
  - Discussion and hands-on work: PAB 360 (“computer lab”)

Flipped Classroom

- "Flipped classroom":
  1. Lecture videos and homework assignments will be posted by Friday evening (YouTube and GitHub)
  2. Monday afternoon, you will be asked to fill out:
     - A short quiz covering the material (Canvas)
     - Anonymous survey about anything that was unclear in the lectures (Canvas)
  3. TTh: We'll spend our in-person time discussing and doing homeworks
     - Group discussion (groups of 2-3), followed by a joint discussion (~40 minutes)
     - Work on the homework (groups of 2-3). Ideally, you can finish your homework in class!

Communication: Slack

- In this class, we won't be using a mailing list but an instant messaging (-like) tool called Slack (http://slack.com). Slack is heavily used today by many research & technology companies and projects.
- Signing up for Slack:
  a) https://join.slack.com/t/uw-astronomy/signup
  b) Join the #astr-324 channel

- What to use it for:
  - Asking questions, discussing the class, exchanging snippets of code, discussing homeworks.
  - Please prefer asking questins via Slack to sending me e-mails. Two reasons:
    - Everyone can benefit from the question and answer.
    - Your colleagues may be able to help!

Action: Let's make sure everyone's on Slack!

Course Materials

- I'll be adding most of what we need to the following organization on GitHub:

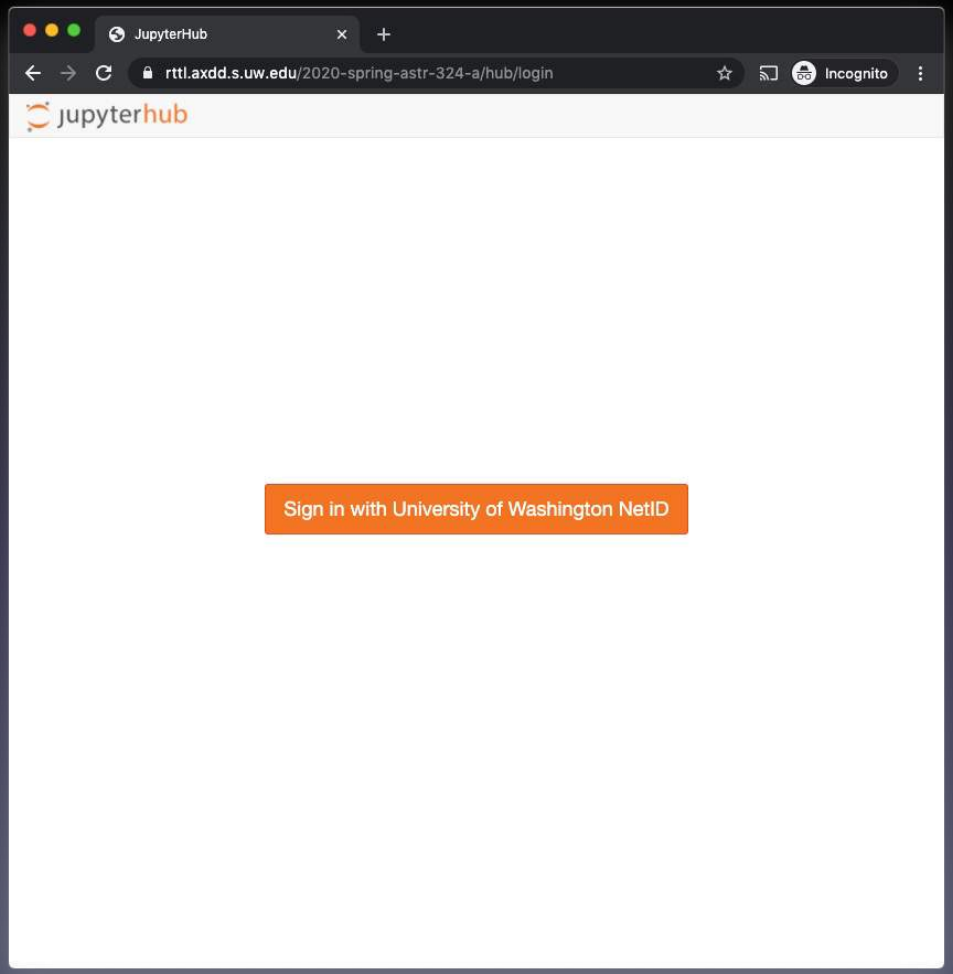
  https://github.com/uw-astr-324

Action: Let's make sure everyone can access this repository.

JupyerHub

- JupyterHub: running Jupyter notebooks remotely.
- Go to:

  https://dirac.us/hub324
- This environment has all you need for the class. No need to install anything on your laptops!
- Caveats
  - To conserve resources, it will close your notebooks after 1hr of inactivity.



Action: Let's try this out (breakouts)!

Homeworks and Grades

Homeworks (70% of the grade):

- Jupyter notebooks. Designed to exercise what we've learned in any given week. Roughly ≤one per week.
- All homeworks will be turned in via JupyterHub, two weeks after being assigned.
- Grading: Will drop your lowest scoring homework (missed homeworks count as zero pct.). Will add +10% for homeworks turned in within 1wk (next Friday).
- Late homework policy:
  - -20% for being up to 1wk late
  - -50% for more than 1wk late.

Final exam (20% of the grade):

- Largely simple questions with few-sentence answers asking about the key concepts we've discussed in class. Very similar to weekly quizzes.
- Will be "take home", closed book & limited time, on an honor system (administered via Canvas).

Quizes (10% of the grade):

- Multiple-choice questions every due every Monday afternoon (5pm).

Prerequisite: Python

- This class will focus on statistics and machine learning; we will not have time to teach coding. I will assume you have a working knowledge of Python programming at the level of ASTR 300
- This includes:
  - Elementary Python, including:
    - understanding built-in Python objects (lists, dictionaries, strings, files)
    - flow control statements (if statement)
    - for and while loops
    - writing and calling functions
  - Familiarity with common Python modules
    - numpy, matplotlib
- **If you're rusty at any of this, please brush up on your Python skills over the next week or two.**

All This and More: Syllabus

- https://github.com/uw-astr-324/astr-324-s22/blob/master/syllabus/syllabus.pdf

ASTR 324: Introduction to Astrostatistics and Machine Learning in Astronomy

Mario Jurić

University of Washington, Spring Quarter 2022

Location and Time: TTh 10:00am-11:20am, PAB 360

Office Hours: After Thursday class

Grading: homeworks, 70%; final exam: 20%; quizes: 10%.

Class materials: `https://github.com/uw-astr-324/astr-324-s22`

Class JupyterHub: `https://dirac.us/hub324`

UW Astronomy Slack: `https://join.slack.com/t/uw-astronomy/signup`, then join `#astr324`

Textbook: Ivezić, Connolly, VanderPlas & Gray: *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*

Flipped classroom with online teaching:
This course will follow the flipped classroom model. In this method of teaching, you will listen to (prerecorded) lectures at home, and come to class (virtually, via Zoom) to engage in discussion, group work, and work on homeworks.

Learning statistics and ML for the first time is far from easy, but it's certainly doable!

For example, this class in 2019:

- Minimum grade: 3.2
- Mean: 3.8
- Median: 3.9

"It is said that despite its many glaring (and occasionally fatal) inaccuracies, the Hitchhiker's Guide to the Galaxy itself has outsold the Encyclopedia Galactica because it is slightly cheaper, and because it has the words 'DON'T PANIC' in large, friendly letters on the cover."

Douglas Adams, The Hitchhiker's Guide to the Galaxy