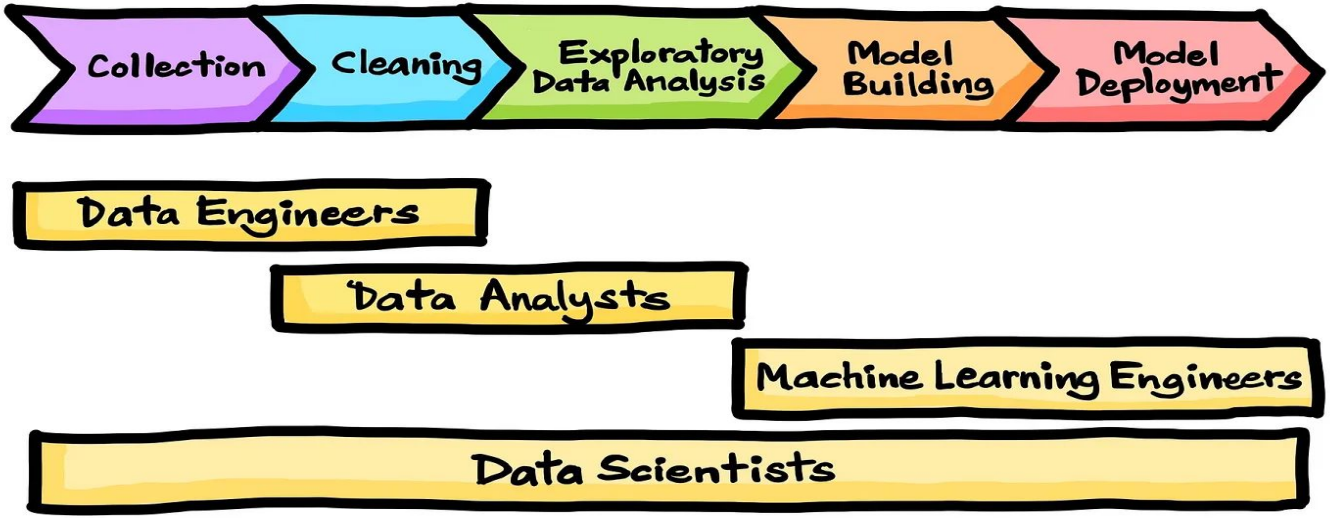
COGS 108 - Final thoughts

Key takeaways

- **Data Wrangling:** Have enough data points (multiple data sources may help), and don’t forget to get permission from author (if applicable), always cite your sources.
- **Data Cleaning:** Always get rid of null/NaN rows, drop irrelevant columns.
- **Data-preprocessing:** Make sure to process data as per your model requirements, for ex: converting column to date_time, getting full_name column from first_name and surname, normalization, etc.
- **EDA:** Make sure to add meaningful visualizations, start with basic ones and then move onto more complex patterns.
- **Machine Learning:** Split data for training, validation and testing. Make sure to train on sufficient data, and stop training once validation accuracy starts falling down.
- **Evaluation:** Use proper metrics to evaluate your model on test data: accuracy %, f1-score, IoU, etc.

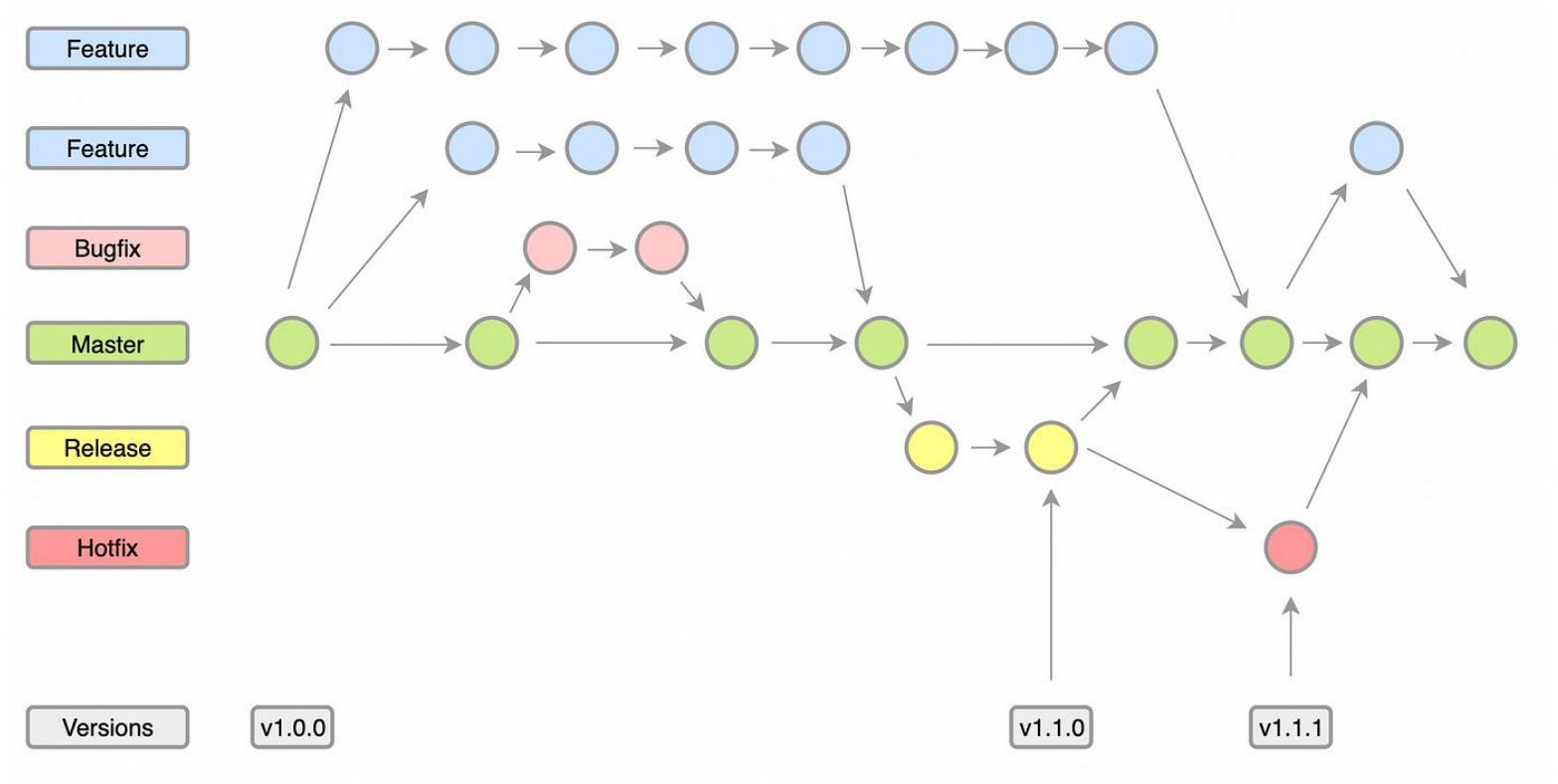
THE DATA SCIENCE PROCESS



Git - Why is it important to learn?

- Imagine a MNC with its vast array of software projects and a large number of developers spread across different locations.
  - Git's distributed nature allows each team to work independently on their local repositories and merge changes seamlessly when needed.
  - Git enables code reviews, where changes proposed by developers are reviewed and approved before being merged.
  - Git records every commit with information about the author, making it easy to trace back and identify who made specific changes, enabling developers to take responsibility for their work.
  - Git integrates well with issue tracking systems, making it easier to link code changes to specific issues, enhancing bug fixing and resolution processes.

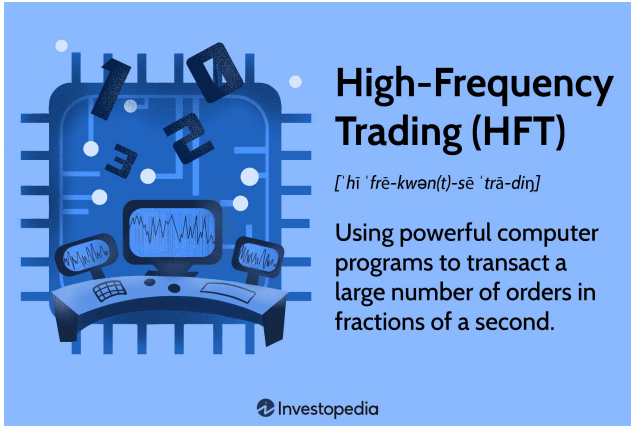
Git workflow in industry



https://betterprogramming.pub/a-simple-git-repository-strategy-93a0c7450f23

Life as a Data Scientist at a High Frequency Trading Firm

- Working at an HFT firm as a data scientist means being part of a fast-paced and dynamic environment.
- Data scientists are constantly analyzing and processing large volumes of financial data in real-time to make quick and informed decisions.
- Data scientists develop and optimize algorithms to predict market movements and identify trading opportunities with high accuracy.
- Working in HFT can be stressful but exciting as well, as decisions need to be made quickly and accurately.



Importance of Online Resources

- When **learning** data science in an academic or professional setting, it is important to utilize online resources
- APIs, Google, and the internet are your best friends
- Things are constantly changing - very less is static when it comes to this field, so it is crucial to keep learning and updating yourself
- Having resources open to refer to while learning is beneficial - don’t put pressure on yourself to memorize everything
- Experience and practice are the best teachers
  - https://scikit-learn.org/stable/index.html
  - https://pandas.pydata.org/
  - https://numpy.org/doc/stable/index.html

Data-Related Opportunities on Campus for Undergrads

- Research labs - data analysis with biological, scientific, social, psychological, and more types of data
  - Opportunities on: Handshake, Real Portal UCSD, AIP Portal UCSD
  - https://real.ucsd.edu/
  - https://ucsd.joinhandshake.com/login
  - https://aip.ucsd.edu/
- Geisel Library - Data & GIS Labs
- Academic Internship Programs (AIP) - quarter or more long, teaching + working combined, usually on-campus
- Data is everywhere!

All the best for Discussion, Assignment and Project submissions. Reach out to us if you need help!
It was great working with you all!