ONLINE MASTERS IN DATA SCIENCE

DSC 208R - Data Management for Analytics

Data Collection and Governance

Arun Kumar

UC San Diego

COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

Review Questions

- Name 3 common forms of structured data with application examples.
- What is Parquet? Explain 1 pro and 1 con of using Parquet vs. CSVs for DS applications.
- What is a data lake? How is it different from an RDBMS?
- Explain 2 reasons why data acquisition can be challenging and how to mitigate them.
- Explain 2 best practices for data reorg/prep for analytics.
- Explain 1 pro and 1 con of programmatic labeling over hand labeling of ML data
- Name a data privacy law that affects many Web companies.
- What is data provenance? Why should it be tracked?