MAS DSE 260: Capstone Project

İlkay ALTINTAŞ, Ph.D.

Lecture 3: Exploratory Data Analysis

Today's Topics

1. Reviewing where we are
2. STEP III: Exploring Data
3. Report III Format : DUE 2/15/18

General Feedback

Report 2

- More focus on success metrics around data transfer, querying, updates and modeling
- Action-oriented steps
- When do I know when to iterate? What are control points?
- A few teams focused on automated workflows, the more the better

Presentation 1

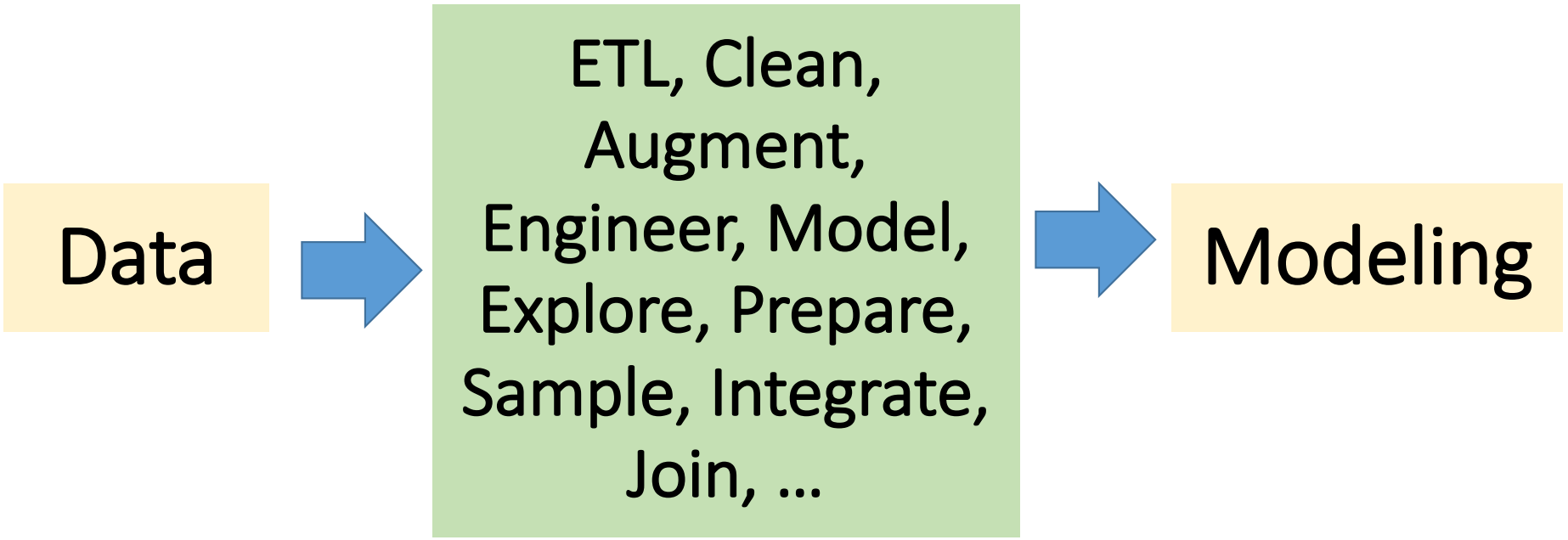
- Remember your imaginary audience
- Explaining the challenge a lot more clearly
- Graphics should be clear and referenced if not original
- Ending on your wins so far, e.g., early EDA results

Process Roadmap (260 A)

- ✓ Step 1: Understanding the Challenge
  - ✓ REPORT 1: due 1/18
- ✓ Step 2: Designing the Data Acquisition and Preparation Pipelines
  - ✓ REPORT 2: due 2/1
- Step 3: Exploring Data
  - ✓ PRESENTATION 1: 2/3
  - REPORT 3: due 2/15
- Step 4: Defining Your Hypothesis and Minimum Viable Modeling Product
  - REPORT 4: due 3/1
- Step 5: Creating a Solution Architecture for Modeling and Optimization
  - PRESENTATION 2: 3/3
  - FINAL WINTER REPORT: due 3/16

Exploratory Data Analysis (EDA) and Pre-Processing

Data Pipelines for EDA



EDA Objectives

- Come up with a clear hypotheses related to the question
- Eliminate/add/clean/augment data
- Evaluate statistical inference of observed trends
- Assess and plan data management and modeling techniques, tools and infrastructure
- Come up with a baseline and strategy for iterations
- Collect metrics for feasibility and scalability requirements in the long term

How do you present EDA progress and results?

- REPORT YOUR INTERPRETATION AND HYPOTHESIS
  - Anything of statistical significance
  - You are trying to understand the data and fix it when needed
  - Most of the activity if not reportable

- FOCUS ON REPRODUCIBILITY
  - Repeatable actions
  - Coe versioning and repositories

- EXPLAIN HOW IT INFLUENCED DATA MODELING AND ENGINEERING

NEXT: Think towards your MVP!

Step III Report Guidelines

- Title, team members and advisor(s)
- Sections:
  - Key Findings through EDA (Different for each project)
  - Data Exploration, Cleaning, Wrangling and Engineering
    - Data Exploration Summary
    - Data Preprocessing
    - Storing processed and/or integrated data
      - Processed dataset description for each processed dataset including why you want to process it that way
      - Table for processed data sets including processed data set name, input datasets, link to the processing scripts and notebooks, and provisional data size
    - Feature Engineering and Data Modeling
      - Summary of feature sets
      - Table for feature set including links to input datasets, feature engineering scripts and notebooks, and provisional data size
    - Approach for Data Access
      - Design for data querying interfaces
      - Justification for manual vs. programmatic access
  - Bullets for each team member's individual contributions in Step 3
  - Any major updates to Steps 1 and 2 as a result of exploratory data analysis
- Keep it to 4-6 pages
- Due date: 2/15/2018 midnight

Questions?

ILkay Altintas, Ph.D.
Email: ialtintas@ucsd.edu