MSDS 694
Distributed Computing

DIANE WOODBRIDGE, PH.D

Your Team

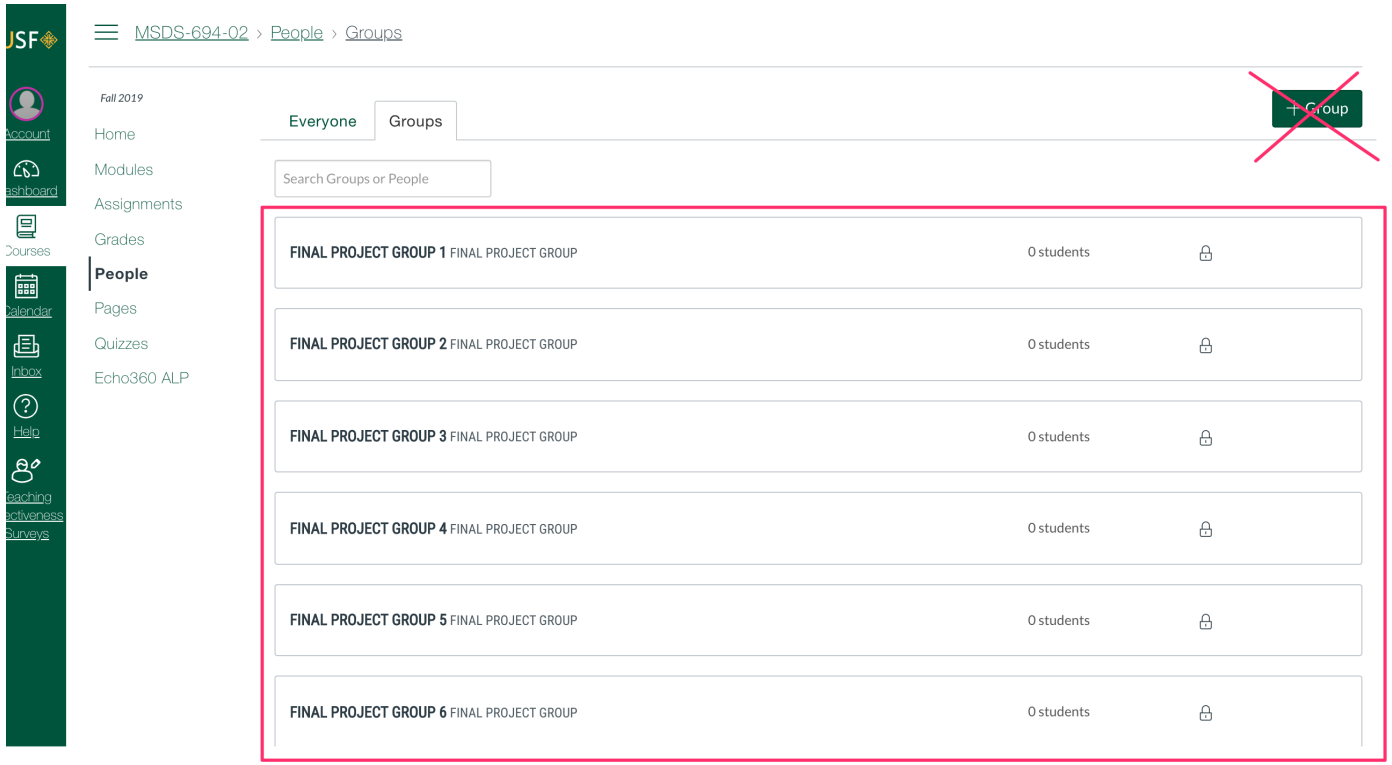
Three to Five people per team.

- Everyone needs to work !! (We are going to survey.)

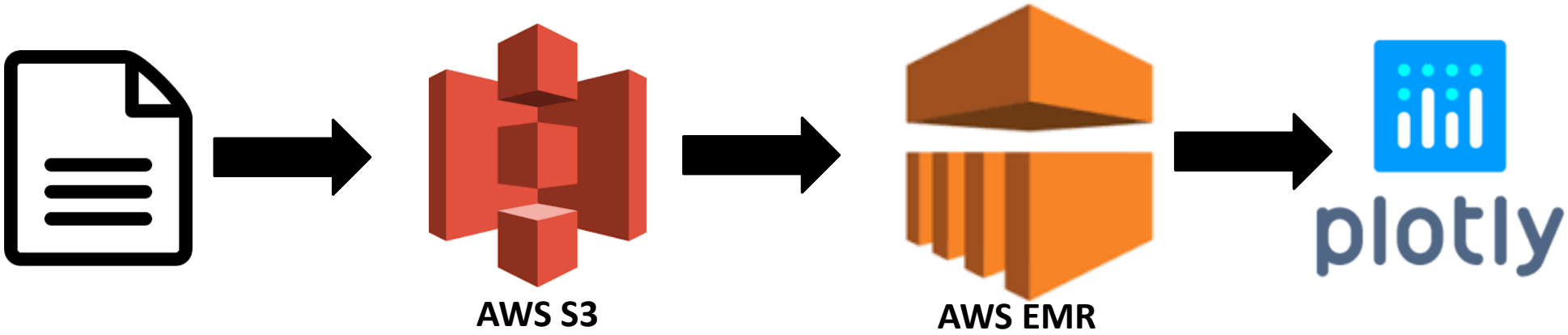
Your Team

Three to Five people per team.

- Everyone needs to join Group under People (on Canvas)



Group Project



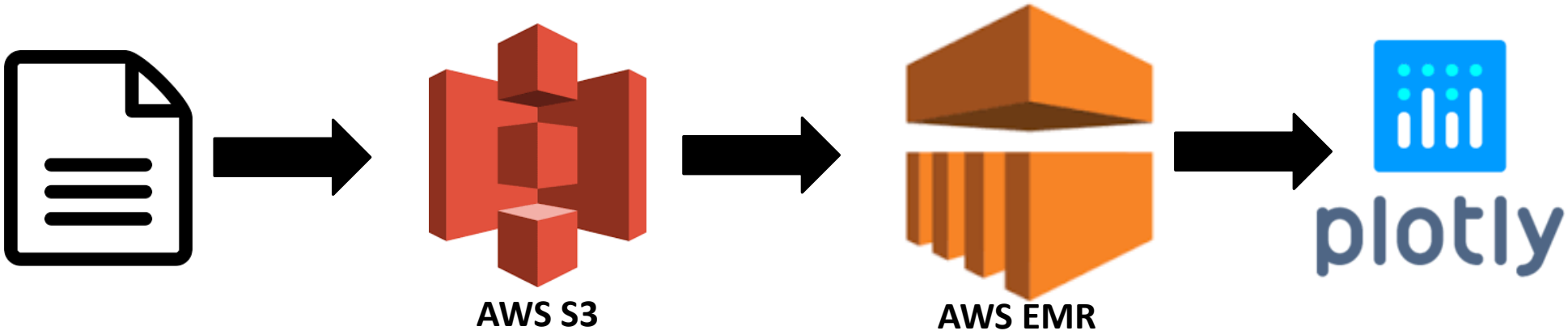
Step 1. Choose Data Sets

Step 2. Load to S3

Step 3. Apply data preprocessing to retrieve interesting stats/trends.

Step 4. Visualize the outcome.

Group Project



Step 1. Choose Data Sets

Step 2. Load to S3

Step 3. Apply data preprocessing to retrieve interesting stats/trends.

Step 4. Visualize the outcome.

Which Data?

Some Data Example

Awesome Public Data : https://github.com/awesomedata/awesome-public-datasets

Kaggle Data Sets : https://www.kaggle.com/datasets

UC Irvine Machine Learning Repository : https://archive.ics.uci.edu/ml/index.php

Data.gov : https://www.data.gov/

Registry of Open Data on AWS : https://registry.opendata.aws/

Requirements

Submit a 1-page data description.

1. **Each group member** proposes one or two topics (3pt).
   - List Student Name, Data titles, List data sources (URLs), Size, Reasons why you chose, Possible analytic goals.
     - If you are planning to collect your own data (web crawling, smartphone application, IoT application, etc.), please describe your specific plans and timeline.
   - Data should be at least 2GB and over 1M records.
2. **Each group** chooses one data set (2pt).
   - Describe reasons why you chose.

Things to Be Considered

Novel data analytics goals

- Data fusion from multiple data sources
- Develop(or Apply) and validate novel algorithms
- Compare results of different algorithms
- Compare results from different machine specs- Costs/Speed

Questions?

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**