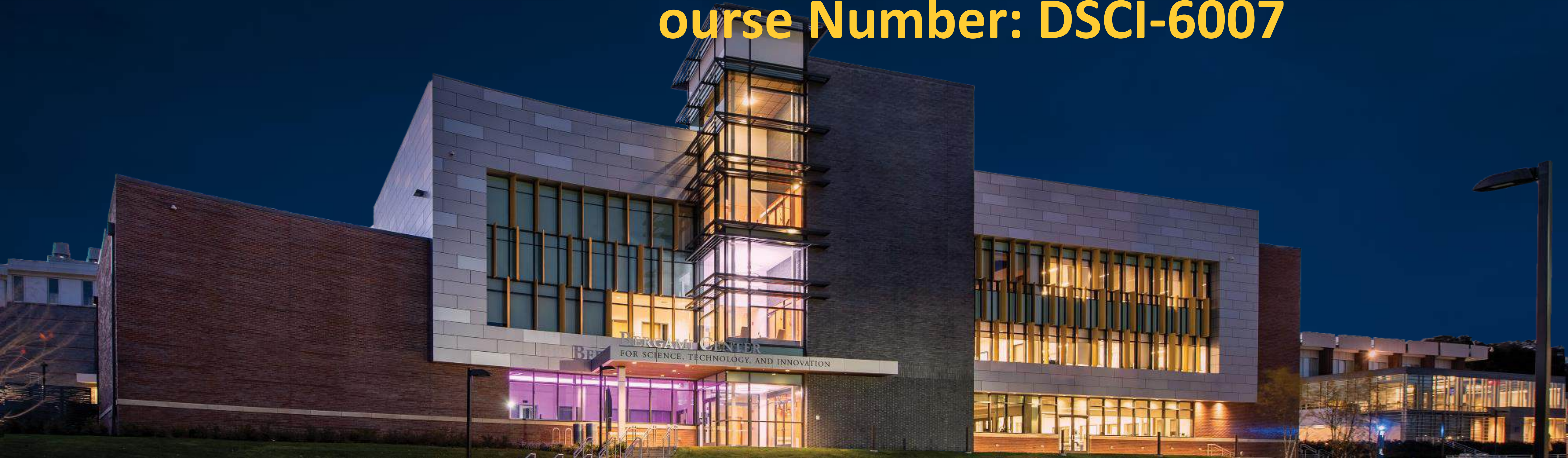
University *of* New Haven

TAGLIATELA COLLEGE OF ENGINEERING

POWER ON

NEWS ARTICLE DATA PIPELINE



ourse Number: DSCI-6007

TEAM-11

Thomas Powell

Team Leader
Data Scientist

Simran Sattar

Data Analyst

Pavan Kumar Sunkara

Data Engineer

University of New Haven

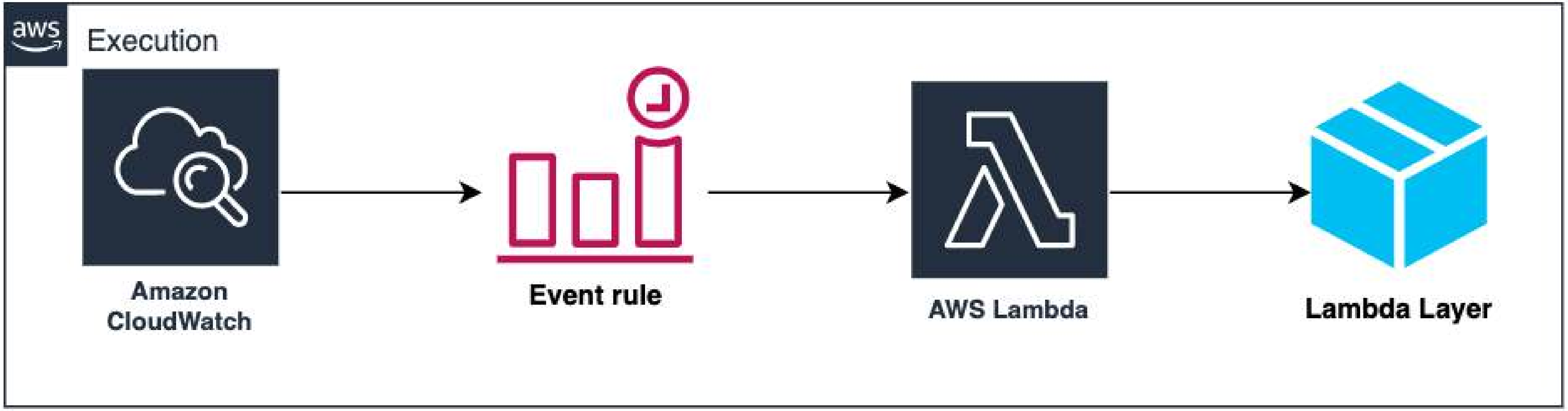
Business Scenario Overview

- Problem
  - It is extremely difficult to keep track of current events with the 24-hours news cycle and multitude of platforms that publish news.
  - It is impossible to deduce the personal importance or relevance of stories
- Data Needed
  - News Articles
    - Headlines, Media Outlet, Publishing Date
- Data Obtainment
  - Google News webpage

Solution Overview

- Business Understanding
  - Keep track of most popular topics, trends in daily news.
- Data Understanding
  - Scrape headlines from Google News articles.
- Data Preparation
  - Serverless Architecture (AWS Lambda), Web Scraping (BeautifulSoup), HTTP requests (Requests library), Data Storage (S3), Error Handling (AWS Cloudwatch).
- Modeling
  - Topic Modeling (Latent Dirichlet Allocation)
- Evaluation
  - Topic Coherence, Bias Checking, Article Consistency.
- Deployment
  - Automation with AWS Lambda, Storing the Data with AWS S3, Monitor Lambda execution logs in AWS CloudWatch.

Architecture diagram of the solution


Execution
Amazon CloudWatch
Event rule
AWS Lambda
Lambda Layer