BIG DATA, Section INET

Assignment 3 (15 points) is your final class project ***proposal.***

Due Wednesday, October 27th, 11:55PM.

1. **Project Idea.** What you are proposing to build/experiment/solve, etc. The project MUST use big data technologies, and MUST be a problem that is NOT solvable using a single computer. This does not mean your project cannot run on a laptop, but that MUST be scalable to big data IF it were ever put into a production environment.

2. **Team members:** Teams of 3, or 4 or 5 are OK and encouraged. If you have a bigger team, get my approval first and be ready to justify it. Bigger teams are encouraged, since we only have a certain amount of time for in class presentations.

Project Proposal

PROBLEM STATEMENT

By nearly every metric, the wildfires in the Western United States are worsening. They are growing larger, spreading faster and reaching higher.

--NYT, July 2021

Wildfire poses a severe and increasing threat to our property, our climate, and our lives. In recent years, state and national governments have introduced bills, issued RFPs, and contracted with engineering firms to improve their ability to remediate and suppress fires, with limited effect.

According to NFPA, the US Federal Government has gone from spending $450mn to $1.6bn per year on wildfire suppression between 1985 and 2021.

And yet, this amount is dwarfed by wildfire's cost. Repairing the damage from 2018's wildfire season cost over $100bn. 2020's wildfire season is expected to cost $150bn.

SOLUTION PROPOSAL

We propose using Big Data technology to explore the enormous climate and imagery datasets now publicly available for researchers and extract unique lessons and actionable insights.

Specifically, we will draw on publicly available datasets from NOAA/AWS, USGS and/or NCWG to derive new insights about the recent increase in wildfire frequency, spread, and intensity. Questions we will aim to explore include --

- Why was the 2021 fire season so unusual? (Why are fires spreading faster? Why are they burning hotter? Why are they changing direction and speed more rapidly than in prior years? Why are traditional methods of fire modeling and containment proving less effective than in prior years?)
- What are some unique ways to visualize the human and financial toll of these fires, using big data tools?
- What can we expect in 2022 and beyond, if current trends continue? (Should we expect more? How many? Where? What new regions might become wildfire hotspots as warming continues?)

We will employ a wide array of analysis and visualization tools, including but not limited to Apache Spark, Dask and MLlib. We will containerize the key components of our solution in Docker, and provide an outline for how we would/could scale the solution using Kubernetes.

Team Members

Benjamin Feuer
Dennis Pang
Jinyang Xue
Subei Han
Yuvi Raina

Project Roles

Role	Member(s)	Task	Tech Used	Due Date
HPC “Devops”, data management, repo	Ben Feuer	Create and maintain github repo. Manage data on the HPC cluster. Create dockerfile for container build/access.	(HPC, Linux/unix, github, Docker)	
Data Ingest, Exploratory Analysis: NOAA	Ben Feuer	Figure out best practices for working with data, schema, perform EDA, develop hypotheses	(Colab, Pyspark, Pandas, Dask)	
Data Ingest, Exploratory Analysis: USGS (Pyspark, Pandas, Dask)	Jinyang Xue Dennis Pang	Figure out best practices for working with data, schema, perform EDA, develop hypotheses	(Colab, Pyspark, Pandas, Dask)	
Data Ingest, Exploratory Analysis: NCWG (Pyspark, Pandas, Dask)	Subei Han	Figure out best practices for working with data, schema, perform EDA, develop hypotheses	(Colab, Pyspark, Pandas, Dask)	
Statistical Analysis	Ben Feuer Yuvraj Raina	Use traditional statistical	(Colab, Pyspark,	

(Hypothesis testing, regression analysis, aggregates, etc)		methods to generate insights about the data	Pandas, Dask)	
AI/ML Analysis	Ben Feuer Subei Han	Use deep learning networks to model topics of interest, classify risk levels, etc.	(MLlib, Pytorch)	
Data Visualization	Yuvraj Raina	Generate attractive and original visualizations of the most important points in the data	(Dask, Seaborn, Plotly, Matplotlib, Pandas, BI)	
Final Write-up and submission	Ben Feuer	Summarize and lay out our findings	(Markdown)	