Assignment 1
CSP554 - Big Data Technologies
Nishant Khandhar (A20581012)

1. What was the problem with the Google flu detection algorithm?

Ans: It kept guessing way too high, at one point more than double the CDC's numbers in 2012-13 and it missed big moments like the 2009 H1N1 wave. A big reason: it learned lots of wintery search habits (not just flu searches), and the link between what people searched and real flu cases wasn't stable. Simple models that just used recent CDC data actually did better.

2. What is big data hubris?

Ans: It's the trap of thinking huge datasets can replace careful, trusted data and basic domain knowledge. Big data is powerful, but it can also be noisy and biased, so it should add to, not replace, traditional sources.

3. What approach could have been used to improve the Google flu detection algorithm?

Ans: Combining GFT with traditional surveillance (e.g., lagged CDC data), dynamically recalibrating the model, and applying standard statistical checks would have substantially improved accuracy.
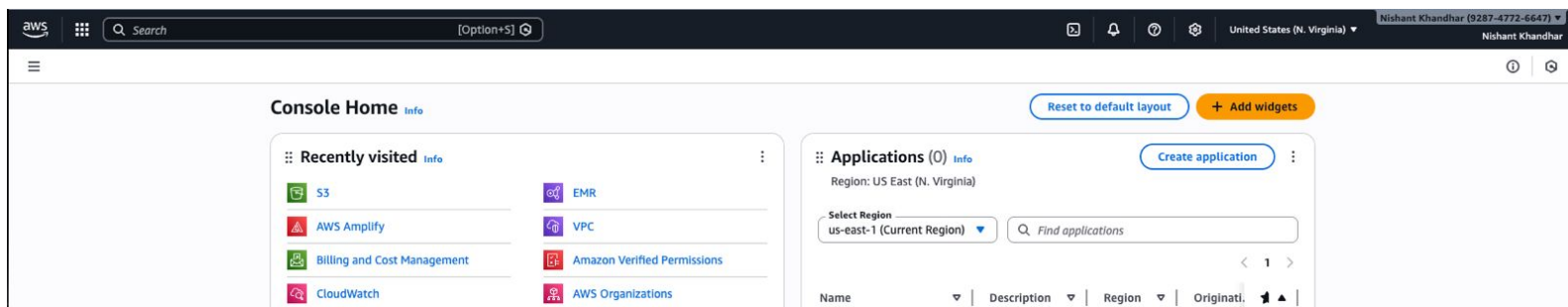
4. What is "algorithm dynamics?"

Ans: Algorithm dynamics are changes over time in the data-generating process caused by platform updates and evolving user behavior, which can destabilize measurements.

5. What aspect of algorithm dynamics impacted the Google flu detection algorithm?

Ans: Frequent changes to Google's search algorithm (e.g., suggested searches and health-related prompts) altered user queries and relative search volumes, breaking the assumed stable link between searches and actual flu prevalence.

- **Screenshot of the main page of the AWS management console page:**

- **Screenshot showing some named object is in the S3 bucket:**