Notes: The Parable of Google Flu — Big Data Pitfalls

Pitfall	Explanation
Overfitting	The model captured noise and coincidental correlations instead of real causal patterns. GFT mistook spikes in “flu symptom” searches as actual outbreaks.
Overparameterization	Too many parameters made the model overly flexible. With thousands of variables and no regularization, GFT fit historical data but failed in prediction.
Lack of Ground Truth Validation	No continuous calibration with CDC data; errors accumulated over time.
Algorithmic Drift	User search behavior and media attention changed, breaking previous correlations.
Correlation ≠ Causation	Search volume increases reflected public concern, not infection rates.
Lack of Transparency	The algorithm and variable list were not publicly released, preventing replication.
Neglecting Traditional Data	Ignoring epidemiological data led to unstable predictions; hybrid models performed better.
Big Data Hubris	The belief that massive data alone can replace theory and domain expertise.