IS883: Deploying Generative AI

Mohannad Elhamod

BOSTON UNIVERSITY

Boston University Questrom School of Business

Financial Modeling of Gen AI Applications

BOSTON
UNIVERSITY

Boston University Questrom School of Business

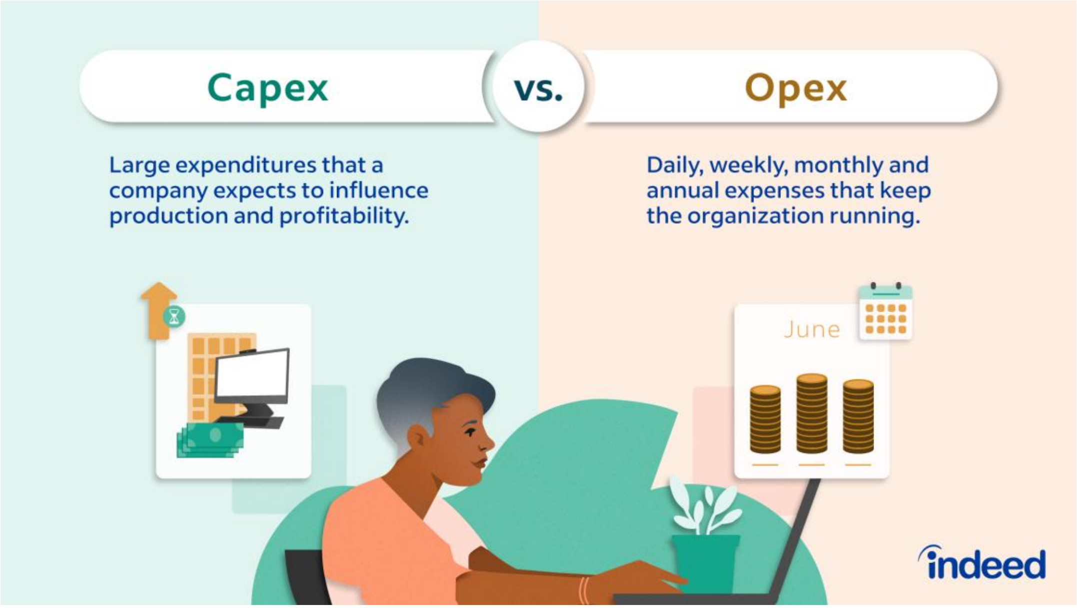
Cost-Benefit Analysis

- When building a Gen AI solution or application, we need to assess its ***viability***.
- We generally would look at the ***ROI*** (Return on Investment)

$$\mathrm{ROI} = \frac{\text{Net Profit}}{\text{Total Cost}} = \frac{\text{Total Revenue} - \text{Total Cost}}{\text{Total Cost}}$$

Total Cost

- Total cost can be broken down into two types:
  - **Upfront Costs (CapEx):** Do not repeat.
  - **Ongoing Costs (OpEx):** Recurring expenses to keep the service running.


Capex
vs.
Opex
Large expenditures that a company expects to influence production and profitability.
Daily, weekly, monthly and annual expenses that keep the organization running.
June
indeed

Upfront Costs

- Examples

```
    isUrl       = ( (type === "url") ||
    isElement   = ( (type === "element") ||
    isObject    = ( (typeof $object ===

// Check if boxer is already active,
if ($("#boxer").length > 1 ||
    return;
}

// Kill event
_killEvent(e);

// Cache internal data
data = $.extend({}, {
    $window: $(window),
    $body: $("body"),
    $target: $target,
    $object: $object,
    visible: false,
    resizeTimer: null,
    touchTimer: null,
    gallery: {
        active: false
```

Software and Tools

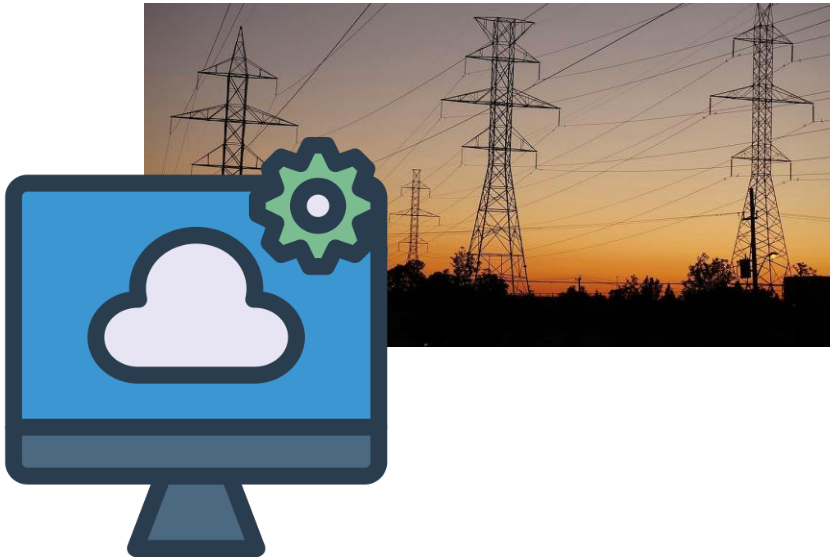
Data

Staffing

Infrastructure

Ongoing Costs

- Examples

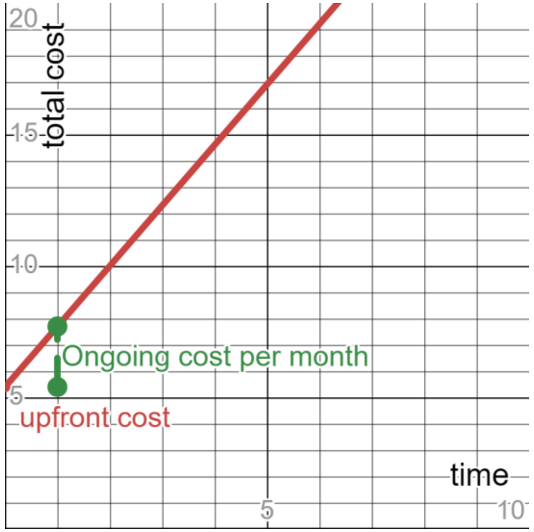


Operational costs (Energy, subscriptions, etc.)

Maintenance costs

Why Care For This Distinction?

- Because that’s what matters for the cost equation.
  - Your decision should be based on:
    - Cashflow: **How much** money you have and **when**.
    - Product Lifespan.
    - Scalability and Flexibility: Pay-as-you-go provides more control and easier adaptation.
    - Risk: Lower upfront commitment is more risk averse.



$$\text{Total Cost} = \text{Upfront Costs} + (\text{Ongoing Cost per Unit of Time} \times \text{Time Period})$$

What Affects Cost in Gen AI?

In-house Approach	Pay-as-you-go Approach
• **Infrastructure** (upfront or ongoing). • **Software licensing** (ongoing) • **Developers and Staff** (upfront and ongoing). • **Maintenance, energy, and upgrades** (ongoing). • Includes model size (computational needs) • **Training cost** (upfront or ongoing): • Data size	• **Developers and Staff** (upfront and ongoing). • **Cloud subscription fees** (ongoing). • **Usages** (ongoing): • Number of tokens (i.e., request and response sizes). • Model type.

Total Revenue

- Total revenue depends on the provided service/product, but could be measured as
  - **Cost Savings:** If product or service is to be used internally, how much cost reduction is there due to automation and increased efficiency?
  - **Sales Increase:** due to an increase in customer satisfaction.

$$\text{Total Revenue} = (\text{Sales per Period} \times \text{Price per Unit} + \text{Cost Savings per Period}) \times \text{Time Period}$$

$$\text{Cost Savings} = (\text{Reduction in Personnel} \times \text{Personnel Cost per Person})$$

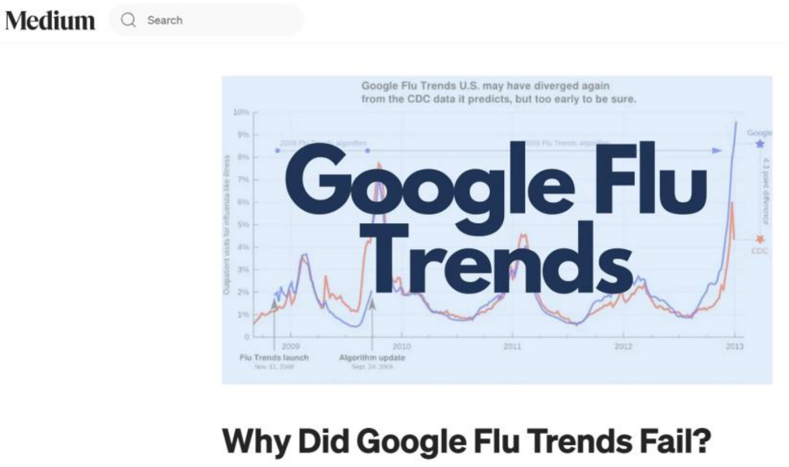
Evaluations: ROI, CBR, and Break-Even Analysis

- Break-Even Point: Time it takes for Total Revenue to become equal Total Cost.
- Cost-Benefit Ratio is simply: $\text{CBR} = \text{ROI} + 1$
  - ROI is negative when product/service is not viable.
  - CBR is less than one when product/service is not viable.

- Try it here!

Uncertainty and Risk

- May not be straight forward to incorporate.
- Attributed to several factors:
  - Regulations and compliance.
  - Long-term performance.


Medium
Search
Google Flu Trends U.S. may have diverged again from the CDC data it predicts, but too early to be sure.
Google Flu Trends
Flu Trends launch
Algorithm update
2009
2010
2011
2012
2013
Google
CDC

Why Did Google Flu Trends Fail?

Risk Evaluation

- We will not discuss this in detail as it requires going into statistics.
- But, in its simplest form, you could account for worst- and best-case scenarios to study the sensitivity of your estimates.

Resources

- Gen AI: too much spend, too little benefit?
- Understanding the Cost of LLMs