FNAL AI Infrastructure Planning: Emerging Technology Directorate Needs

Gabriel Perdue

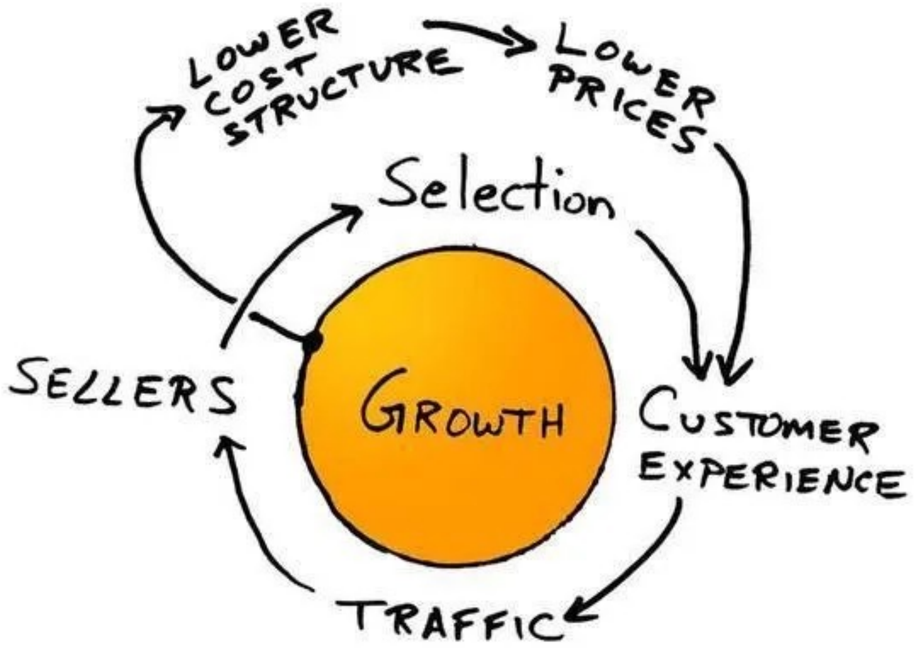
6 / April / 2023

Overview

- Philosophy: Our strengths ARE our weaknesses
- ETD teams look very different than "typical" particle physics experiments.
- Current solutions are not really sustainable
- Need to build for flexibility

Philosophy

- I (we) will proceed with some complaints, but…
- It is worth asking - **how do organizations achieve excellence?**
  - One perspective suggests organizations become very high-performance when all of the pieces are naturally aligned to common goals, and efforts are self-reinforcing and combine coherently to create an output larger than the sum of all the individual parts - ***the flywheel****.

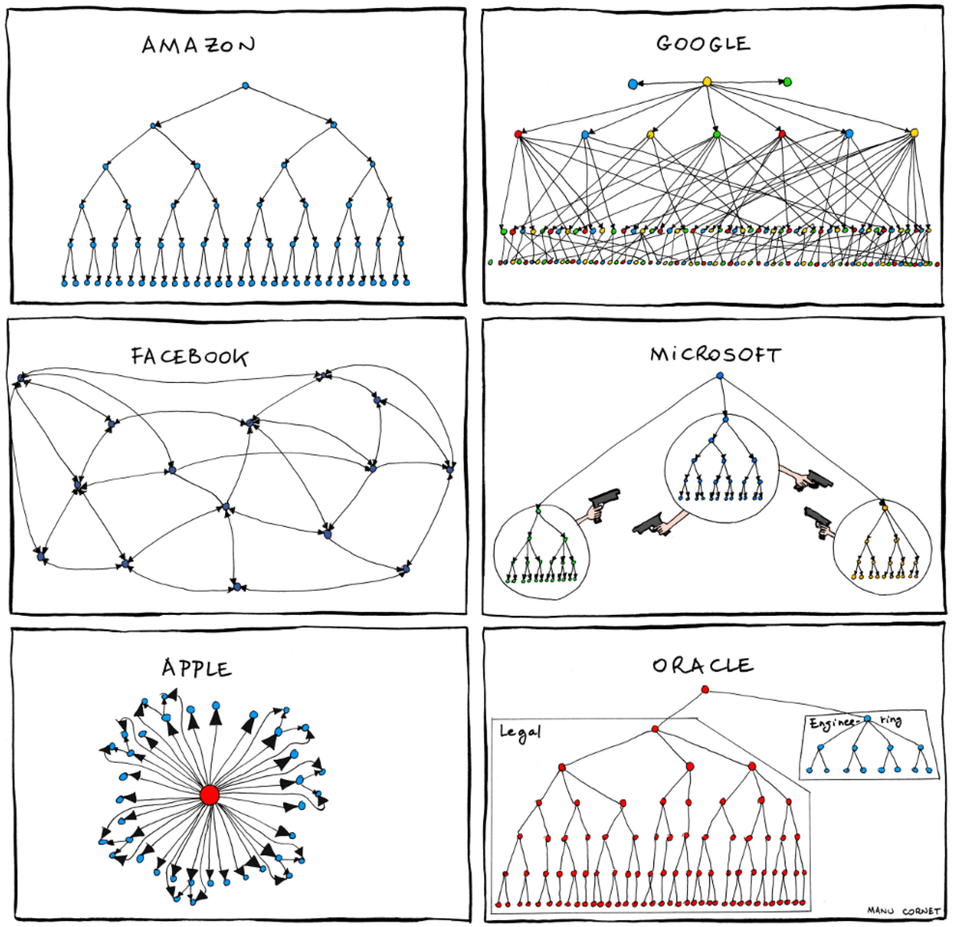


* B. Thompson, https://stratechery.com/2021/the-relentless-jeff-bezos/

Fermilab

More Philosophy!

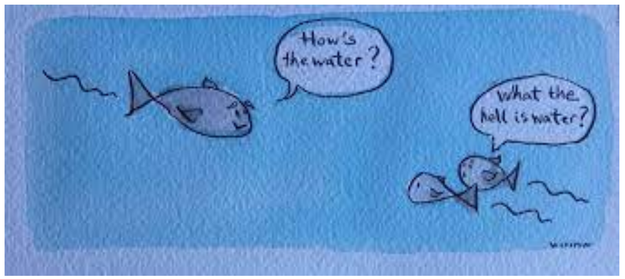
- But... **what happens when you achieve excellence at one task, and then decide you want to do something different?**
  - Think about **why dominant companies eventually fail** when the economy undergoes a fundamental shift - it is **NOT** because they were idiots.
  - This is a real challenge because ***the things that make your organization excellent in one area almost always necessarily make your organization under-perform in a different one***.
- *Many of the computing applications in ETD "suffer" because Fermilab is so well oriented to serving the needs of very different users.*



* https://bonkersworld.net/organizational-charts

What do we do? (for work I mean, in ETD)

- Quantum open system simulation for pulse engineering and optimal control problems.
- Quantum circuit simulation.
- Machine learning *for* quantum simulation problems.
- Data analysis of simulations / materials data.
- Asking what we do with AI is like asking a fish “how is the water?”*



- Most expensive is quantum simulation - exponential scaling costs mean that exact simulation becomes intractable *very* quickly.
- Quantum simulation *sort of* looks like machine learning in the sense that it is driven by linear algebra engines and many platforms that are good for ML are also good for quantum simulation.
- Data analysis looks like ~industry… so scipy, some Julia, etc. - ROOT is completely unheard of…

* https://fs.blog/david-foster-wallace-this-is-water/

Computing teams and tasks in ETD look... different

- ***Teams are smaller***
  - Some range:
    - Typical is one PI with a postdoc and a couple of students. The students may be from institutions without prior history with Fermilab (so, no cooperative agreements in place at the start).
    - Even in larger groups (e.g. SQMS), computing projects generally contain less than 10 people.
  - Teams cohere and decohere quickly - little opportunity for institutional memory or scripting tools to develop.
    - *Documentation must be current and ideally there is a Helpdesk that responds to questions within a day.*
- Tasks look more like HPC problems and less like HTC problems
- Data is messier, and HEP-centric metadata solutions are overkill
- Datasets are also much smaller and the total amount of compute required to to an analysis is much smaller - this makes *the overhead / work ratio very different than even for small HEP experiments*
  - This is true even when thinking about quantum simulation, where it is feasible (though very rare) for very expensive quantum simulations to require compute of order of an HEP experiment (only concentrated into a several-day burst on an LCF).

Current solutions are not sustainable…

- Essentially all of my team's computing problems are solved like this:

Current solutions are not sustainable…

- Essentially all of my team's computing problems are solved like this:

*

(Sorry for the low-res photo Burt!)

Fermilab

*Just ask yourself - has anyone seen Burt and Batman in the same room? Didn't think so…

Need to build for flexibility

- **“We” (small teams) should be careful what we wish for because we might get it.**
- What we do now actually works pretty well in terms of getting results over the past few years:
  - Commercial cloud, with a sprinkle of EAF, coordinated and managed by Burt
- But, there are some **scaling problems**:
  - *Commercial cloud is pay as you go and more expensive than FNAL computing and full of little gotchas that threaten to make it even more expensive. Funding is still mostly through OHEP and they don’t like to budget for cloud compute, so we end up having to cut labor, basically.*
    - Maybe that’s just life.
  - Also, as great as Burt is, it isn’t fair to make him a single point of failure

Need to build for flexibility, cont.

- A very formalized ticket system through SNOW is probably not ideal for provisioning resources, creating accounts, removing access, deleting old cloud storage, etc. either.
- The process super-structure, the Fermilab *flywheel*, that makes large experiments work smoothly will probably not work for small teams because the needs change a lot from project to project and the overhead of adapting may be expensive.
- How to solve this?

  - ***I wish I knew!***
- Maybe something like HEPCloud is the solution. But…
  - Still need support for small teams in terms of setting up accounts and managing resources. Still need a reconceptualized “Helpdesk.” Still need different hardware.