



Amenti K. Kenea (1)

▼ Contact Links

Email	amenti@uni.minerva.edu
Phone	(917)679-7391
Arena/General Visual Taste	https://www.are.na/amenti-kenea/channels
Fashion/Creative Side	https://aava.club/
LinkedIn	https://www.linkedin.com/in/amenti-kenea/

▼ Weekly Reading Roll

Last Updated: "Week Ending": 03-1

Mountain Dew's Twitch AI Raid

- I'm split on how I feel about this. Incredible way of marketing to the right audience by cornering true fans. However, I worry how intrusive this could get.
- Are we entering a new era of affiliate marketing and product placements?
- "During the live period, the RAID AI will crawl all concurrent livestreams tagged under Gaming looking solely for MTN DEW products and logos. Once it identifies the presence of MTN DEW, selected streamers will get a chat asking to opt-in to join the RAID. Once you accept, the RAID AI will keep monitoring your stream for the presence of MTN DEW, if you remove your DEW, you'll be prompted to bring it back on camera, if you don't, you'll be removed from our participating streamers."

▼ Radiooo Project

I BUILT AN ONLINE RADIO THAT I UPDATE EVERY 2 DAYS. I love sharing music w friends and wanted a corner on the internet to do that.

<https://aava.club/songs>

Thesis: there is something to be said about curation in the day we live in. We grew up on it. Music discovery was MTV and the billboard top 100. I was plugged into the radio growing up to just know what was considered "good"/"cool".

Now it's more general public consensus. TikTok allows for true virality of a snippet of a song. Twitter/Instagram/YouTube allow for open conversations on what's good/not. What's hot and what's not.

Current state of streaming and radio: Streaming revolutionised music listening by giving users on-demand access to everything — the opposite of radio. But, over time, streaming

Abstractions Rule Everything Around Me - Benjamin Schneider

- "I realized that people came up with some of the abstractions most impactful in our everyday lives without ever referring to either! The more you notice all the abstractions you interact with, the more coming up with useful abstractions starts to look something humans are just generally interested in — and pretty good at."

Yudkowsky vs Hanson on FOOM: Whose Predictions Were Better? - 1a3orn

- I alternate between worried/excited with all the recent ai this/that debates — esp. around agi or interpretability voids. It was fun looking back at debates in the course of ML over years in the rationalist community and what they got right/wrong. This is a good summary of Eliezer and Hanson's predictions.

Are you serious? - Visakan Veerasamy

- "So the point is to take the work seriously but you don't take *yourself* too seriously. There's a riff about this in Stephen Pressfield's War of Art, where he talks about how amateurs are too precious with their work: '*The professional has learned, however, that too much love can be a bad thing. Too much love can make him choke. The seeming detachment of the professional, the cold-blooded character to his demeanor, is a compensating device to keep him from loving the game so much that he freezes in action.*'"

has started to look a lot like its predecessor. Streaming services now push algorithmically-generated playlists and ready-made mixes to soundtrack activities, like working out and cooking. Spotify's AI-powered voice DJ is a lot like listening to a radio DJ provide context on their curated mix of songs. We even have streaming "stations"! Where is all this heading?

Given the noise, the true winners can be picked. They're generally agreed on. It became democratic.

But it isn't inherently democratic. Sometimes we need the "real ones" or the "cultural curators" to tell us what's good and what isn't.

NOTICE: GenZ didn't know MTV or Radio. What good was decided with consensus instead of up down.

The recent incredible growth of "Youtube Reaction Channels" is an indication of that. Which leads me to...

We need new methods of content recommendation/curation that's based on the curator's taste.

Derrick Gee: he's a TikTok previous radio show host that has very respectable and professions (but still loving) insight on music. People started flooding into hearing what he proposed. He sort of started becoming a tastemaker for people that wanted to escape the current musical bubble.

He started playlists (including other Spotify playlist makers that became professional discovery helpers).

This was the initial trigger for an inspiration that I've believed for a while. it's not

- "I'm still publishing. That's the litmus test. Are you publishing, whatever publishing means to you? I want to see it!"

Resignation Letter - Sam Hinkie

- clarity, brevity, and specificity in summarizing his objectives
- "A competitive league like the NBA necessitates a zig while our competitors comfortably zag. We often chose not to defend ourselves against much of the criticism, largely in an effort to stay true to the ideal of having the longest view in the room."

Why Generative AI Is Mostly A Bad VC Bet - Rob May

- Surprisingly early (Jan 7) call on why LLM Startups might not be the move. + I like Rob

When the cost of something trends towards zero because of new technology:

1. You will get an explosion of that good.
2. That good will decline in value and defensibility
3. The economic complements to that good that see increased demand as a result of the explosion in the original good, will be the place to invest.

THE NEXT ACT OF THE GVASALIA BROTHERS CIRCUS: Eugene Rabkin

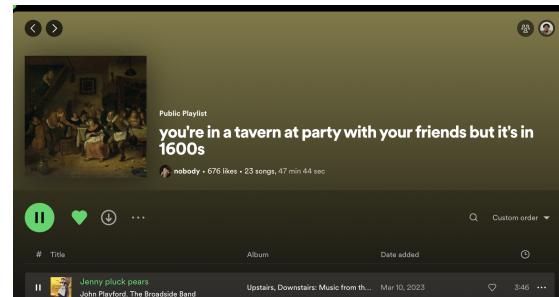
- "It sounds bizarre, like a desperate couture attempt at streetwear, or worse, like a Marie Antoinette playing-at-shepherdess scenario."

"This is just the latest chapter in the Gvasalia circus, which, sadly, the

anything that's new.

One thing that I know I'm good at is galvanizing a direction so that people buy shit that's cool.

- Taste + Momentum + Leading



New twist on online listening:

- Poolsuite (one oliver sent)
- Radiooooo
- <https://danielpmarks.github.io/vinyl-stream/>
- <https://www.last.fm/dashboard>

True Radio (just online format that allows for discovery)

- Radio Garden – Ethio FM 107.8

Radio that took true online form

- The Lot Radio
- Dave & Central Cee pass through the booth for a special episode of Victory Lap | Balamii
- Lower Grand Radio

Personally, it feels very rewarding as it feels like the intersection of all the things that I like: Imagine being Marty – the founder of poolsuite – and saying this. This is exactly all of my worlds colliding.

fashion commentariat cannot get enough of."

a Nirav or a Naval - Auren Hoffman

It's very important to realize what you're changing or chasing. You have the ability to revolutionize a bunch of things as you're deffo an outsider. Never discredit that. And don't let the fact that you sometimes appear as an insider to gain clout, make you inherently an insider that's un-opinionated/dull/and unable to influence a tectonic change.

Superliner Returns - Paul Graham

- "always be learning. If you're not learning, you're probably not on a path that leads to superlinear returns."

Why Do Rich People In Movies Seem So Fake? - Michella Jia

- "If you are excellent in the first way, it behooves you to control the contexts in which you perform — and if you can control these contexts well, you also come off well. As for the second form of excellence, it often appears latent until catastrophe or circumstance forces a change of context. In this sense, the second type of excellence is much more difficult to spot."

Telomeres: Everything You Always Wanted To Know - Nintil

- "The usual function ascribed to telomeres is as an anti-cancer mechanism: if we cell begins dividing too much then its telomeres will progressively shorten and it will stop dividing (or die). To overcome this,



marty.jpg • [Follow](#)
Stanly Ranch, Napa, California



marty.jpg Honoured to have spoken at @bof's first Global Beauty Forum alongside legends @isamayaffrench & @imranamed, telling the story of how we built good-will via @poolsuite w/ 8 yrs of free internet music experiences before launching our spin out sunscreen co @vacationinc 🌟 which you can now find in >2,500 retailers across the US incl. @kith @ultabeauty @anthropologie @nordstrom @thestandard @acehotel @urbanoutfitters & more! PS. We got to hang with Pamela Anderson and she's REALLY LOVELY

I think there is something to be said about the original style (retro desktop) vibe that poolsuite.fm had created.

It could be vinyl, cassette, or CD players. Or it can randomly simulate other stuff. I think it's incredible.

One thing about this musical experience is I want it to be as fun as humanly possible and interpreted on a computer as it can possibly become. FUN AND CULTURED.

The theory goes: limited (curation mattered but it was solo dumping, not essential) → universal consensus voting (multiple options and choose based on actual democracy) → age of too many good options (tell me what I like)

<https://www.are.na/amenti-kenea/radio-iqgpifzfzno>

▼ don't walk in soho

a run club for meeting people and connecting with young professionals in the city

Instagram: [don't walk in soho](#)

Website: [don't walk in soho](#)

cancers end up reactivating telomerase to keep their telomere length."

An Extremely Opinionated Annotated List of My Favorite Mechanistic Interpretability Papers - Neel Nanda

- "The core thing to take away from it is the perspective of networks having legible(-ish) internal representations of features, and that these may be connected up into interpretable circuits. The key is that this is a mindset for thinking about networks *in general*, and all the discussion of image circuits is just grounding in concrete examples. On a deeper level, understanding why these are important and non-trivial claims about neural networks, and their implications."

▼ aa-va fashion

aava.club

<https://aava.club/>

<https://www.are.na/amenti-kenea/aa-va>

▼ Database View (Status and Dates Written)

Amenti's Internet Stuff

Date	Name	Status	Category
@2024/01/20	⌚ Weekly Reading Roll	Reviewing	MISC
@2023/10/27	📺 Phoebe Philo	Done	Fashion

Date	Name	Status	Category
@2023/11/15	 J Hus	Done	Music
@2022/02/03	 Lending and Default Likelihood Prediction	Done	Statistics
@2023/04/17	 Early AI Meditations 3	Live	Writing
@2023/04/05	 Early AI Meditations 2	Live	Writing
@2023/03/27	 Early AI Meditations 1	Live	Writing
@2023/11/21	 Radiooo Project	In Dev	MISC
@2023/12/28	 Beyond Prompts	Reviewing	Writing
@2023/02/15	 Neural Networks from Scratch	Done	Statistics
@2022/02/25	 Assessing the Influence of Neutral Grounds on Match Outcomes	Done	Statistics
@2022/03/08	 Case Against Bloated MVPs	In Dev	Writing
@2022/10/20	 Improving Series A Working Capital	Live	Writing
@2023/07/04	 Vector Embeddings - Hype from Excess Dry Power?	In Dev	Writing
@2023/08/08	 Suddenly Popular LLMOps	Reviewing	Writing
@2023/06/05	 Glenn Martens - Vanguard of the Modern Silhouette?	Live	Fashion
@2023/10/24	 Paris Spring '24 Men's Quick Reflections	Live	Fashion
@2022/04/07	 Approaching Ideal Money - Are Non-Collateralized Algorithmic Stablecoins Feasible?	Done	Statistics
@2024/03/13	 Anomaly Detection in Timeseries Data	Live	Statistics
	 Task	Not started	
	 Task	Not started	

Tech

▼ Simpler Alternative to X-Learner for Uplift Modeling

A guide to the simplified X-learner approach for Uplift modeling

Meta-learners like S-Learner, T-Learner, and X-Learner are some of the most widely used approaches for Uplift modeling. When teaching about these approaches, I find that students often find the X-learner model somewhat confusing to understand. In this post, I describe a modified approach I call simplified X-learner (Xs-learner) that is easier to understand, faster to implement, and in my experience often works as well or better in practice.

Uplift Modeling

A/B testing is a common method used at tech companies to make informed decisions. For example, imagine you want to send out a coupon to users and you want to know how much it will increase the chances of them completing their first order with your service. By running an A/B test, you can determine on average how effective the coupon is. However, you may also want to know which users the coupon will help you generate higher profits and which users the coupon will cause you to lose money.

Uplift modeling is a technique that lets us go beyond learning the average effect of a treatment and instead helps us understand how the effect of the treatment varies across your users. This allows us to more efficiently decide which treatment to send to each user.

Meta-learners

Some of the most common approaches for solving uplift problems are known as meta-learners, because they are ways to take existing supervised learning algorithms and using their predictions in order to make estimates of the treatment effect for each user.

I'll be demonstrating each of these approaches using a dataset from Lenta, a large Russian grocery store that sent out text messages to their users and saw whether it would increase their probability of making a purchase. In each of the examples I will be using the following notation:

- Y: Did the user make a purchase (the outcome variable)
- T: Did the user receive a text message(the treatment variable)
- X: All the other information we know about the user, e.g. age, gender, purchase history. (The Lenta dataset has almost 200 features describing each user)

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from xgboost import XGBClassifier, XGBRegressor
from sklift.datasets import fetch_lenta
from sklift.viz import plot_qini_curve
from numpy.random import default_rng
rng = default_rng()
```

We'll use the sklift package, which has a useful function that helps download the data for the

Lenta uplift experiment and do some basic processing of the data.

```
data = fetch_lenta()
Y = data['target_name']
X = data['feature_names']
df = pd.concat([data['target'], data['treatment'], data['data']], axis=1)
gender_map = {'K': 0, 'M': 1}
group_map = {'test': 1, 'control': 0}
df['gender'] = df['gender'].map(gender_map)
df['treatment'] = df['group'].map(group_map)
T = 'treatment'# Split our data into a training and an evaluation sample
df_train, df_test =
train_test_split(df, test_size=0.3, random_state=42)
```

S-Learner

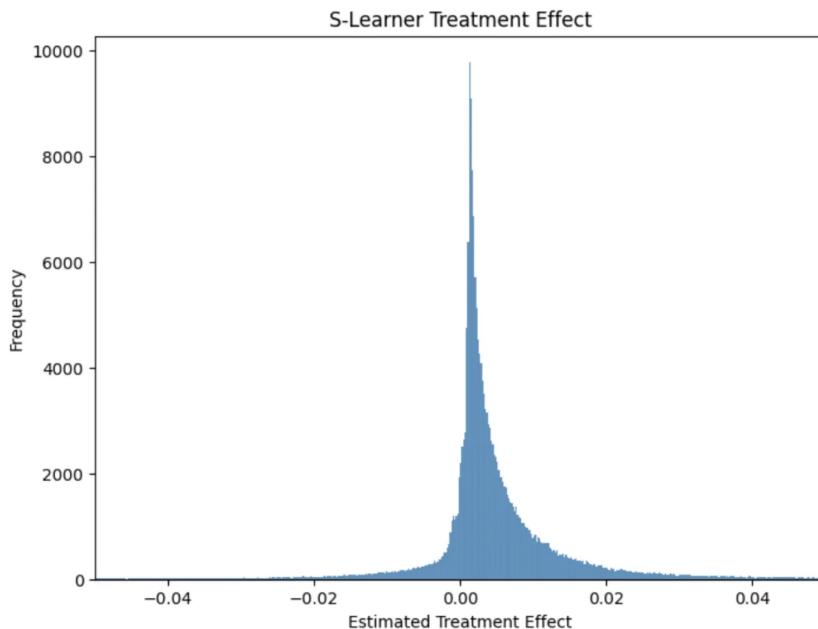
S-learner is the simplest and easiest to understand of these approaches. With S-learner you fit a single machine learning model using all of your data, with the treatment variable (did you get a text message) as one of the features. You can then use this model to predict “what would happen if the user got the text” and “what would happen if the user did not get the text”. The difference between these two predictions is your estimate of the treatment effect of the text message on the user.

In all my examples, I use XGBoost as a simple and effective baseline ML model that is fast to train and generally works well on many problems. In any real world problem you should be testing more than one type of model and should be doing cross validation to find hyperparameters that work well for your particular problem.

```
slearner = XGBClassifier()
slearner.fit(df_train[X+[T]], df_train[Y])
# Calculate the difference in predictions when T=1 vs T=0# This is our estimate of the effect of the coupon
for each user in our dataslearner_te = slearner.predict_proba(df_test[X].assign(**{T: 1}))[ :, 1] \
- slearner.predict_proba(df_test[X].assign(**{T: 0}))[ :, 1]
```

One downside of the S-learner model is that there is nothing that tells the model to give special attention to the treatment variable. This means that often your machine learning model will focus on other variables that are stronger predictors of the outcome and end up ignoring the effect of the treatment. This means that on average your estimates of the treatment will be biased towards 0.S-learner treatment effect distribution

The mean treatment effect is 0.005.



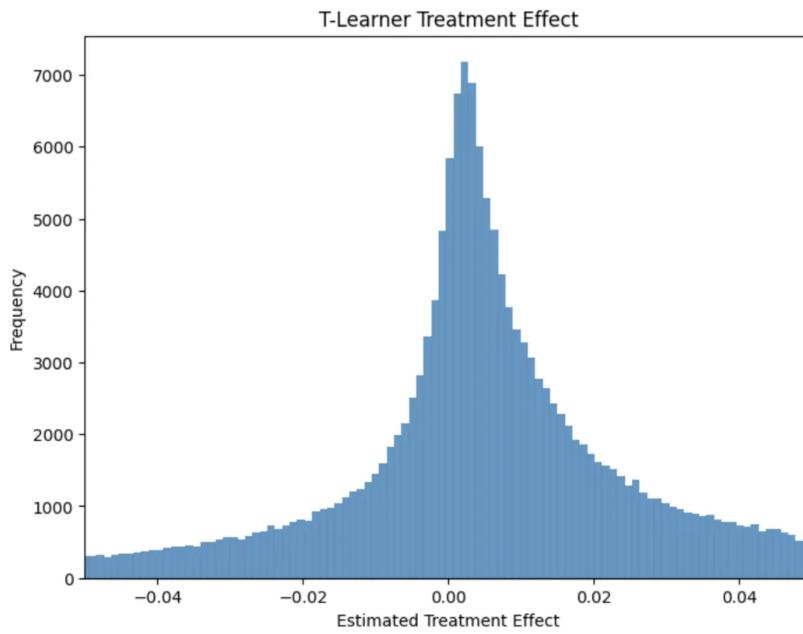
S-learner treatment effect distribution

T-learner

T-learner uses two separate models. The first model looks only at the users who did not receive the coupon. The second model looks only at the users who did receive the coupon. To predict the treatment effect, we take the difference between the predictions of these two models. T-learner essentially forces your models to pay attention to the treatment variable since you make sure that each of the models only focuses on either the treated or untreated observations in your data.

```
tlearner_0 = XGBClassifier()
tlearner_1 = XGBClassifier()
# Split data into treated and untreated
df_train_0 = df_train[df_train[T] == 0]
df_train_1 = df_train[df_train[T] == 1]
# Fit the models on each sample
tlearner_0.fit(df_train_0[X], df_train_0[Y])
tlearner_1.fit(df_train_1[X], df_train_1[Y])
# Calculate the difference in predictions
tlearner_te = tlearner_1.predict_proba[df_test[X]](:, 1) \
- tlearner_0.predict_proba[df_test[X]](:, 1)
```

```
The mean treatment effect is 0.008.
```



T-learner treatment effect distribution

Simplified X-learner (Xs-learner)

The simplified X-learner use 3 models to form its predictions. The first two are exactly the same models we used for T-learner: one model trained only using the treated observations, and the other model trained using only the untreated observations.

With T-learner we formed our treatment effect estimates by taking the difference between the predictions of these two models (predicted outcome when treated minus predicted outcome when untreated). The Xs-learner takes the actual outcome of the user under the treatment their received and compares that to the predicted outcome if they received the other treatment (actual outcome minus predicted outcome).

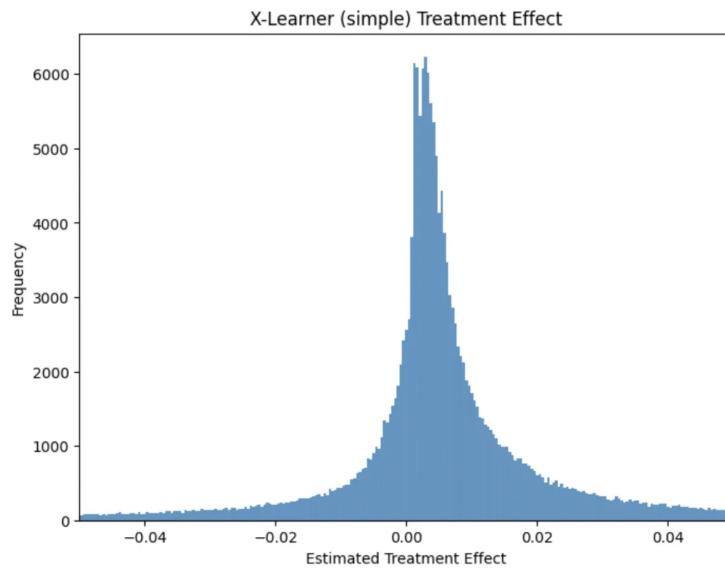
```
# We could also just reuse the models we made for the T-learner
xlearner_0 = XGBClassifier()
xlearner_1 = XGBClassifier()
# Split data into treated and untreated
df_train_0 = df_train[df_train[T] == 0]
df_train_1 = df_train[df_train[T] == 1]
# Fit the models on each sample
xlearner_0.fit(df_train_0[X], df_train_0[Y])
xlearner_1.fit(df_train_1[X], df_train_1[Y])
# Calculate the difference between actual outcomes and predictions
xlearner_te_0 = xlearner_1.predict_proba(df_train_0[X])[:, 1] - df_train_0[Y]
xlearner_te_1 = df_train_1[Y] - xlearner_0.predict_proba(df_train_1[X])[:, 1]
```

We can't use these differences directly, because we would not be able to make predictions for any new users since we wouldn't know the actual outcomes for these new users. So we need to train one more model. This model predicts the treatment effect as a function of the X

variables.

```
# Even though the outcome is binary, the treatment effects are continuousxlearner_combined = XGBRegressor()
# Fit the combined modelxlearner_combined.fit(
    # Stack the X variables for the treated and untreated users pd.concat([df_train_0, df_train_1])(X),
    # Stack the X-learner treatment effects for treated and untreated users pd.concat([xlearner_te_0,
xlearner_te_1])
)
# Predict treatment effects for each userxlearner_simple_te = xlearner_combined.predict(df_test[X])
```

The mean treatment effect is 0.008.



Full X-learner

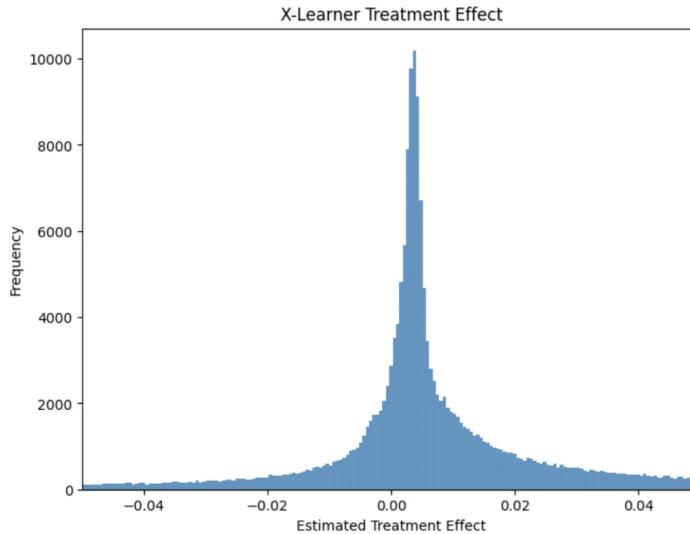
The simplified X-Learner required 3 ML models. The full X-learner as originally proposed by Künzel et al. requires 5 ML models.

Instead of fitting one combined model that predicts the treatment effects for everyone, the full X-learner uses two separate models, one for the treated users and one for the untreated users. This gives us two difference models that can predict treatment effects for new users. Künzel et al. recommend taking a weighted average of the two models, with the weights determined by a final propensity score model that predicts the probability of receiving the treatment.

```
# Define the new models that are not used in the simple versionxlearner_te_model_0 = XGBRegressor()
xlearner_te_model_1 = XGBRegressor()
xlearner_propensity = XBCClassifier()
xlearner_te_model_0.fit(df_train_0[X], xlearner_te_0)
xlearner_te_model_1.fit(df_train_1[X], xlearner_te_1)
# Calculate predictions from both modelsxlearner_te_model_0_te = xlearner_te_model_0.predict(df_test[X])
xlearner_te_model_1_te = xlearner_te_model_1.predict(df_test[X])
```

```
# Calculate the propensity scores
xlearner_propensity.fit(df_train[X], df_train[T])
xlearner_propensities = xlearner_propensity.predict_proba(df_test[X])[:, 1]
# Calculate the treatment effects as propensity weighted average
xlearner_te = xlearner_propensities * xlearner_te_model_0_te + (1 - xlearner_propensities) * xlearner_te_model_1_te
```

The mean treatment effect is 0.0064.



X-learner treatment effect distribution

Comparing the Results

We can compare the performance of each of these models using our held-out test set data. Here I am using Qini plots, which are a common approach for comparing the performance of Uplift models. Similar to an ROC curve, the higher the model's line goes above the diagonal, the better the performance.

```
fig, ax = plt.subplots(figsize=(20, 10))
def plot_qini_short(model, label, color, linestyle):
    plot_qini_curve(df_test[Y], model, df_test[T], name=label, ax=ax, perfect=False, color=color,
    linestyle=linestyle)
plot_qini_short(slearner_te, 'Slearner', 'blue', 'solid')
plot_qini_short(tlearner_te, 'Tlearner', 'red', 'solid')
plot_qini_short(xlearner_simple_te, 'Xlearner Simple', 'purple', 'solid')
plot_qini_short(xlearner_te, 'Xlearner', 'green', 'solid')
ax.legend(loc='lower right');
```

For this particular dataset, the simplified X-Learner had the best overall performance. We shouldn't draw any strong conclusions about the relative performance of different algorithms from this single example. In my experience, which algorithm works best varies a lot

depending on the specific problem you are working on. However, I do think that this example demonstrates that the simplified X-Learner (Xs-learner) is one more approach worth considering when working on Uplift problems.

References

- Athey, Susan, and Guido W. Imbens. Machine learning for estimating heterogeneous causal effects. №3350. 2015.
<https://www.gsb.stanford.edu/faculty-research/working-papers/machine-learning-estimating-heterogeneous-causal-effects>
- Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." Proceedings of the national academy of sciences 116.10 (2019): 4156–4165.
<http://sekhon.berkeley.edu/papers/x-learner.pdf>
- Gutierrez, Pierre, and Jean-Yves Gérardy. "Causal inference and uplift modelling: A review of the literature." International conference on predictive applications and APIs. PMLR, 2017.
<http://proceedings.mlr.press/v67/gutierrez17a/gutierrez17a.pdf>
 - Y: Did the user make a purchase (the outcome variable)
 - T: Did the user receive a text message(the treatment variable)
 - X: All the other information we know about the user, e.g. age, gender, purchase history. (The Lenta dataset has almost 200 features describing each user)
- Athey, Susan, and Guido W. Imbens. Machine learning for estimating heterogeneous causal effects. №3350. 2015. <https://www.gsb.stanford.edu/faculty-research/working-papers/machine-learning-estimating-heterogeneous-causal-effects>
- Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." Proceedings of the national academy of sciences 116.10 (2019): 4156–4165. <http://sekhon.berkeley.edu/papers/x-learner.pdf>
- Gutierrez, Pierre, and Jean-Yves Gérardy. "Causal inference and uplift modelling: A review of the literature." International conference on predictive applications and APIs. PMLR, 2017. <http://proceedings.mlr.press/v67/gutierrez17a/gutierrez17a.pdf>

▼ LLM using Substrate for Robust SQL

Purpose: To build a system that generates syntactically and contextually correct SQL queries from natural language inputs.

This is my experiment to play around with <https://www.substrate.run/> that reduces the complexity of multi-model systems by supporting a graph SDK.

Why do I love substrate?

- I think there should be a platform that takes open source models, speed them up relentlessly, make an api, have the cheapest possible price, great uptime.

- In the long-run, you'll build economies of scale w GPUs + have a process power of optimization.
- There are a very large number of people who want this and at high api volumes right now. It's dead simple. All execution. Performance doesn't just lead to reduced prices, it can lead to better experiences due to less latency.
- Eventually you could intentionally train less powerful models specifically trading off its power for cost/latency to counter position the foundation model focused companies focused on capability more than cost/perf.
- Counterpoint:
 - The worry however is, will this last? I assume the large model builders will quickly get commoditized. Tens of startups will do the same competing for the same \$/developer.
 - Any optimization you do will likely be surpassed by the community. It would be like creating custom optimized version of PHP in 2001. There was a lot of opportunity for performance improvements but no business case here.

Regardless let's see the implementation

Writing SQL with LLMs causes multiple hallucinations not necessarily due to SQL generation, but due misuse in context.

Now that we have bigger context windows, the problem is even worse as dumping all the rows and context to prompts consumes excessive tokens for the simplest queries.

The idea is to find a combination of `Syntax` and `Context` that's robust, but also efficient.

This is done through:

1. mapping of the table being used
2. giving NLP style SQL objects to combine for syntax

Setting Up the Environment

First, let's set up our development environment by installing the necessary Python packages. We'll be using Pydantic for data validation and schema definition. Also let's import the required libraries:

```
pip install pydantic
from pydantic import BaseModel, Field
from typing import Optional, Union, List
from enum import Enum
```

Defining Column Types and Enumerations

We'll start by defining enumerations for our database columns and SQL operations. This will help us restrict the possible values and ensure type safety.

We'll start by defining enumerations for our database columns and SQL operations. This will help us restrict the possible values and ensure type safety.

Setting Up the Environment

```
pip install pydantic
from pydantic import BaseModel, Field
from typing import Optional, Union, List
from enum import Enum
```

Defining Column Types and Enumerations

We'll start by defining enumerations for our database columns and SQL operations. This will help us restrict the possible values and ensure type safety.

```
class Departments(str, Enum):
    IT = "IT"
    SALES = "SALES"
    ACCOUNTING = "ACCOUNTING"
    CEO = "CEO"

class EmpLevel(str, Enum):
    JUNIOR = "JUNIOR"
    SEMISENIOR = "SEMISENIOR"
    SENIOR = "SENIOR"

class column_names(str, Enum):
    EMPLOYEE_ID = "employee_id"
    FIRST_NAME = "first_name"
    LAST_NAME = "last_name"
    DEPT_ID = "dept_id"
    MANAGER_ID = "manager_id"
    SALARY = "salary"
    EXPERTISE = "expertise"

class TableColumns(BaseModel):
    employee_id: Optional[int] = Field(None, title="Employee ID", description="The ID of the employee")
        first_name: Optional[str] = Field(None, title="First Name", description="The first name of the employee")
        last_name: Optional[str] = Field(None, title="Last Name", description="The last name of the employee")
```

```

        on="The last name of the employee")
    dept_id: Optional[Departments] = Field(None, title="Department ID",
description="The department ID of the employee")
    manager_id: Optional[int] = Field(None, title="Manager ID", description="The ID of the manager")
    salary: Optional[int] = Field(None, title="Salary", description="The salary of the employee")
    expertise: Optional[EmpLevel] = Field(None, title="Expertise Level", description="The expertise level of the employee")

```

Defining SQL Syntax Models

Next, we'll define models to describe SQL operations, comparisons, logic operators, and ordering.

```

class sql_type(str, Enum):
    SELECT = "SELECT"
    INSERT = "INSERT"
    UPDATE = "UPDATE"
    DELETE = "DELETE"

class sql_compare(str, Enum):
    EQUAL = "="
    NOT_EQUAL = "!="
    GREATER = ">"
    LESS = "<"
    GREATER_EQUAL = ">="
    LESS_EQUAL = "<="

class sql_logic_operator(str, Enum):
    AND = "AND"
    OR = "OR"

class sql_order(str, Enum):
    ASC = "ASC"
    DESC = "DESC"

class sql_comparison(BaseModel):
    column: column_names = Field(..., title="Table Column", description="Column in the Table")
    compare: sql_compare = Field(..., title="Comparison Operator", description="Comparison Operator")
    value: Union[str, Departments, EmpLevel] = Field(..., title="Value")

```

```

e", description="Value to Compare")

class sql_logic_condition(BaseModel):
    logic: sql_logic_operator = Field(..., title="Logic Operator", description="Logic Operator")
    comparison: sql_comparison = Field(..., title="Comparison", description="Comparison")

class SQLQuery(BaseModel):
    sql: sql_type = Field(..., title="SQL Type", description="SQL Type")
    columns: list[column_names] = Field(..., title="Columns", description="Columns to Select")
    table: str = Field(..., title="Table", description="Table Name")
    conditions: List[sql_logic_condition] = Field(..., title="Conditions", description="List of Conditions with Logic")
    order: Optional[sql_order] = Field(None, title="Order", description="Order")
    limit: Optional[int] = Field(None, title="Limit", description="Limit")

```

Generating SQL Query Structure

Now that we have our models set up, we can use a language model to generate the SQL query structure based on a natural language input. In this example, we'll use OpenAI's GPT-3.5 model, but you could use any suitable language model API.

Let's create a function to generate the SQL query structure:

```

pip install openai
import openai
import json

openai.api_key = 'your-api-key-here'

def generate_sql_json(question: str) -> dict:
    prompt = f"""
        Generate a JSON structure for an SQL query based on the following question:
        {question}

        Use the following JSON schema:
        {json.dumps(SQLQuery.model_json_schema(), indent=2)}

```

```

Respond only with the JSON structure, nothing else.
"""

response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",
    messages=[
        {"role": "system", "content": "You are a helpful assistant that generates SQL query structures in JSON format."},
        {"role": "user", "content": prompt}
    ]
)

return json.loads(response.choices[0].message['content'])

# Use the function
question = "Can you provide me with the amount of employee id and salary in the Account department that has a salary greater than 50000 in descending order?"
json_response = generate_sql_json(question)

# Parse and validate the response
query_formatted = SQLQuery(**json_response)

```

Formatting the SQL Query

Finally, we'll create a function to format the SQLQuery object into a proper SQL string:

```

def format_sql_query(query: SQLQuery) -> str:
    # Generate the initial Base Query with no comparisons
    generated_query = f"{query.sql} {'.'.join([col.value for col in query.columns])} FROM {query.table}"

    # Check for additional conditions
    if query.conditions:
        # Replace first logical operator with WHERE
        generated_query += " WHERE "

        # For each condition, append it the query in the correct format
        for i, condition in enumerate(query.conditions):
            if i > 0:
                generated_query += f" {condition.logic} "
            generated_query += f"{condition.comparison.column} {conditi

```

```

        on.comparison.compare} '{condition.comparison.value}'"

        # if there is an ordering rule, then format and append
        if query.order:
            generated_query += f" ORDER BY {', '.join([col.value for col in
query.columns])} {query.order}"

        # if there is a limit, then format and append
        if query.limit:
            generated_query += f" LIMIT {query.limit}"

    return generated_query

# Generate the final SQL query
final_query = format_sql_query(query_formatted)
print(final_query)

```

This system allows us to generate SQL queries from natural language inputs in a structured and type-safe manner. By using Pydantic models, we ensure that our generated queries adhere to the correct format and data types.

Remember to handle potential errors, such as invalid inputs or API failures, in a production environment. You might also want to add more complex query capabilities, such as JOINs or nested queries, depending on your specific needs.

▼ Retrofitting Access

It blows my mind that people can't project future AI progress onto existing workflows.

Tools like Cursor are already *crazy* useful.

Now apply the next frontier of models, 10M+ token context windows, 1M token output, triple the tokens/sec, etc.

We're just getting started.

▼ LLM for Any Website

I played around with building a python based Streamlit app that lets you chat to any website through RAGs

<https://replit.com/@AmentiKumera/websitechatter#main.py>

▼ Are AI Eval Pseudo-Science?

Automated evaluation platform for LLMs

- Evaluation run to measure model performance
- Adversarial testing sets made to break models
- Generate new adversarial test sets
- Benchmarking performance to other models

Checking if they break in the real world scenario towards regulated industries where there is little tolerance for errors.

- "a trusted third party when it comes to evaluating models."

Rigorous Evaluation

Score + Test + Benchmark

What are my thoughts on the issue being solved:

- LLMs are nondeterministic — they're not guaranteed to produce the same output every time for the same input.
 - That means that companies will need to do more rigorous testing to make sure they're operating correctly, not going off-topic, and providing reliable results.
 - generate adversarial test cases, monitor hallucinations, and detect PII and other unexpected and unsafe behavior. Customers use Patronus AI to detect LLM mistakes at scale and deploy AI products safely and confidently.
- LLMs in current form won't generate any revenue. Enterprises need much more tech & domain training to align to business.
 - Good thing from our enterprise clients that they **are willing to pay** for solution as long as it aligns to business goals

Sideways – test normal modes

- high-priority harm areas: self-harm, physical harm, illegal items, fraud, and child abuse.

Questions

- What are traditional methods of testing
 - academic benchmarks
 - human evaluations
- Why would someone use us instead of public datasets?

There's obviously a clear problem that needs fixing but I worry that it's a hyper competitive space. I'd love to hear about how they're trying to make themselves distinct by directly working with observable surfaces (mongo/hugging face...)

I was looking at Arize and others that seemed slightly more advanced but playing directly on the same field.

From what I can tell, they're relying on the fact that LMs are nondeterministic — they're not guaranteed to produce the same output every time for the same input. That means that companies will need to do more rigorous testing to make sure they're operating correctly, not going off-topic, and providing reliable results. Basically Score + Test + Benchmark... Good thing is they're targeting enterprise clients that are **willing to pay** for a solution as long as it aligns to business goals. Eventually become "a trusted third party when it comes to evaluating models." There's obviously a clear problem that needs fixing but I worry that it's a hyper competitive space. I'd love to hear about how they're trying to make themselves distinct by directly working with observable surfaces (mongo/hugging face...) I was looking at Arize and others that seemed slightly more advanced but playing directly on the same field. What patronus seems to be building is the key features in a stack of a bunch of different processes that are required from a company starting to prep data to getting the final results. Obv it's too early, and I'm in no authority, but I'd assume most of this would be done in one platform that can plug and play different things and the platforms that facilitate that (similar to how you can plug and play with the base models themselves) will eventually build out an evaluation tool. And if there's a better evaluator, it becomes a plug that can be added more than its own platform. Whether they even plan on moving to the entire space of monitor, finetune, observe, test, evaluate or just master the evaluate section and eventually be an option for the evaluation as they're the best... (I'd also assume the best plug-and-play evaluators would become super niche industry specific) Nonetheless I would love to hear from him and am interested in talking about the position you mentioned and getting to learn more about how they're trying to position themselves.

Evals & Leaderboards - what's ultimately the goal?

What is the goal for evaluation and leaderboards?

- Starting from goals for the outcomes we're seeing to have a nuance overview

There are a number of ways to evaluate a model (coding, reasoning, etc..) but the line of sight from something like MMLU or Hellaswag to what the downstream consumer (i.e. the developer) wants in terms of application performance is unclear and certainly non-linear.

. Having evals that accurately represent success in the application environment gives model developers something to shoot for - and god knows, when we get a good benchmark to optimize toward, we as a community go nuts!!

Moreover, while it's still very early, there are signs that we can more confidently start from desired capabilities and adjust the data mixture at various stages of the model development process to align our models more toward how we want them to behave in the eventual application. Whether we need to go all the way back to pre-training is still a good question but we certainly know we can do a lot at the CPT and SFT stages to get models to behave in our desired ways (e.g. speak languages other than english).

Back to goals - what are the goals of evals and leaderboards?

The answer is clearly dependent on who you ask. It can be a combination of wanting to find the best model across performance, type (i.e. base, chat, multimodal, different languages, etc..), latency and cost - or certainly a local optima across all of them depending on who the developer is and what his or her application goal is, level of sophistication is, desired level of customization, etc etc...

As a model developer (or at least one that likes to play with the models out there) I tend to use [Helm](#) or [OpenLLMLeaderboard](#), sort on the provided metrics, see which models rise to the top and which can fit onto a Colab instance so I can just quickly load something up and hack a prototype. My job to be done (JTBD) or goal is pretty straightforward. But for many others, not so much..

Let's walk through some of the leaderboards, some old some new, and what a goal could be for each. Note this is not exhaustive and is a curated set of leaderboards from the articles I read with my personal perspective on each.

(1) Open LLM Leaderboard or HELM

Maintainer: Hugging Face & Stanford respectively

Summary: A sortable set of models, open and closed, evaluated on mostly academic benchmarks. Covers core scenarios such as Q&A, MMLU (Massive Multitask Language Understanding), MATH, GSM8K (Grade School Math), LegalBench, MedQA, WMT 2014 along with other benchmarks. These are largely open source evals that have been developed over the years by various researchers.

Developer goal: General AI developers looking for models to build on will come here first, understand per a given size, which models perform the best and then leverage the result for their research, an application, fine tune for a custom application. One thing that's clear however is that there is a stage of disillusionment that the community is in when it comes to **actual** performance of models and how each rank on academic evals. This has driven the development of other leaderboards such as Chatbot Arena (see later).

(2) Hallucinations Leaderboard

Maintainer: Hugging Face

Summary: It evaluates the propensity for hallucination in LLMs across a diverse array of tasks, including Closed-book Open-domain QA, Summarization, Reading Comprehension, Instruction Following, Fact-Checking, Hallucination Detection, and Self-Consistency. The evaluation encompasses a wide range of datasets such as NQ Open, TriviaQA, TruthfulQA, XSum, CNN/DM, RACE, SQuADv2, MemoTrap, IFEval, FEVER, FaithDial, True-False, HalluEval, and SelfCheckGPT, offering a comprehensive assessment of each model's performance in generating accurate and contextually relevant content.

Developer goal: Have a starting point to understand which models hallucinate the least. There is some overlap with other available leaderboards here and, given without RAG or search

integrated, these models will **always** hallucinate, this leaderboard is less interesting for developers.

(3) Chatbot Arena / LMSys

Maintainer: Together AI, UC-Berkeley, Stanford and several others.

Summary: A benchmark platform for large language models (LLMs) that features anonymous, randomized battles in a crowdsourced manner. Some, like [Karpathy](#), find it to be the best test of a model's performance as the arena is much more dynamic and compares models head to head.

Developer goal: Understand how models will perform in a **real** environment that challenges in ways that status datasets / prompts can't. Like #1 above, this may influence the developer's starting point for further work.

(4) The MTEB leaderboard

Maintainer: Hugging Face & Cohere

Summary: MTEB or "Massive Text Embedding Benchmark" contains 8 embedding tasks covering a total of 58 datasets and 112 languages. Through the benchmarking of 33 models on MTEB, establish the most comprehensive benchmark of text embeddings to date.

Developer goal: Understand which models could perform best in applications where retrieval augmented generation (RAG) is used.

(5) Artificial Analysis

Maintainer: Independent / Startup

Summary: Artificial Analysis provides benchmarks & information to support developers, customers, researchers, and other users of AI models to make informed decisions in choosing: Which AI model to use for a given task, and Which hosting provider to use to access the model.

Developer goal: Understand cost, quality and speed tradeoffs across models and model providers (e.g. OpenAI, Microsoft Azure, Together.ai, Mistral, Google, Anthropic, Amazon Bedrock, Perplexity, Fireworks, Lepton, and Deepinfra.)

(6) Martian's Provider Leaderboard

Maintainer: [Martian](#) (startup)

Summary: Martian's provider leaderboard collects metrics daily and tracks them over time to evaluate the performance of LLM inference providers on common LLMs. You can filter and sort that data based on the criteria for your use case.

Developer goal: Understand quickly cost vs. rate limits vs. throughput for a variety of LLM providers and optimize for his or her application.

(7) Enterprise Scenarios leaderboard

Maintainer: Hugging Face

Summary: Enterprise Scenarios leaderboard evaluates the performance of language models on real-world enterprise use cases (at present 6 benchmarks - Finance, Legal, etc..).

Developer goal: This leaderboard is still very much nascent as you can see from the lack of public models competing. Overall though this set of evals would provide developers with an idea of which base models to use as a starting point for a given task. Again very very early.

(😎 ToolBench

Maintainer: SambaNova Systems

Summary: A tool manipulation benchmark consisting of diverse software tools for real-world tasks.

Developer goal: Illustrate which available models perform best for action generation as described in natural language. Note it seems like this project is also a great resource for broader developers for how to implement a particular tool usage (i.e. plug-in) for their model.

Key takeaways:

1. We are in that stage of the market where it's an absolute gold rush - if an entity controls the underlying benchmark, **this can influence the direction of the market**. This is why startups that are dependent on performance / LLM efficiency are pushing on their own benchmarks. If these benchmarks take hold, it will force competitors to follow.
2. We are absolutely in the early innings and nothing has been decided. If you can dream of a use case or requirement, you could probably define an eval and leaderboard.
3. Right now Meta is a passive player and, beyond the CybersecEval we released late last year, ****just**** leverages the currently available evals and leaderboards.

Lastly, as Alan Kay the PARC researcher said "The best way to predict the future is to invent it.." - we are now on the path to AGI and it's time for ****us**** to define the future. Buckle up!!

Cheers!

How do you evaluate models (coding, reasoning, etc...)

- MMLU or Hellaswag to downstream consumer (ie. the developer) wants in terms of application performance.

You can just look at the provided metrics, see which models rank higher and use the one that goes into your code easily to prototype something. Eg:

- Holistic Evaluation of Language Models (HELM)
- Open LLM Leaderboard - a Hugging Face Space by HuggingFaceH4

What's the true need of evaluations?

- Evals are the new PRD?

1. Find the best model across performance types (chat, multimodal, switching languages, etc...)
2. Finding the optimal range for latency and cost
3. Comparing levels of sophistication or customization levels

Having evals that **accurately represent success** in the application environment gives model developers something to shoot for

- When we get a good benchmark to optimize toward
 - without a really strong POV on where we want to be going, it is impossible to get there.
 - Outside of the common ones like cost, safety, quality [case evals], and latency.
- An important one of those challenges relates to how we can reasonably evaluate the capabilities of these apparently highly-skilled models. How do we quantify what they can and cannot do? How can we measure how biased they are, or what is the risk they will express toxic or harmful language? And what are useful automatic signals that can be used to improve them even further?
- We can **work backwards from desired capabilities** and adjust the data mixture at various stages of the model development process to align our models towards how we want them to behave in the eventual application
 - Do we even need to go all the way back to pre-training or not is a different question, but we know that there's a lot we can do even at the CPT and SFT stages to make our models behave in our desired ways... (change languages etc...)
- IMO, eval is the **largest challenge we face**. If we can't show that the models are working/achieving our goals for them, **what are we even doing?** The crux of the issue is that it's a challenging thing to scope out what a "general purpose" system ought to be able to do

Learnings from Current Instances (which are listed in detail below)

- If an entity controls the underlying benchmark, this can influence the direction of the market. Every startup whose product is dependent on performance/LLM efficiency is pushing on its own benchmarks. Once a benchmark becomes the norm competitors are forced to follow. Thus it's a **goldrush to control the underlying benchmark**.
 - The best way to predict the future is to invent it...
 -
- It's **very early** and nothing has been decided. So if you can see a desired capability or requirement, you can define an evaluation and leaderboard.
 - Every LLM org is going to have to define their version of AGI and build the evals towards that. Eg. OAI is biased towards a single agent, Google is benching on "search" for general intelligence.

- Evaluating a model taken to an extreme limit **is as complex as coming up with an intelligent model**. Back in the days (lol) we got away w easier evals bc the dimensionality and decision of taste were very tractable and were often deterministic.
 - With the incredible advances in foundation models to conduct a wide range of tasks, it becomes more and more challenging to evaluate them and understand their limitations: **if a model could potentially execute infinitely many tasks, it is impossible to do a proper evaluation of all the tasks it could do.**
- **Evaluations are subject to human preferences** which vary in small populations. So more models we have, differentiation is preference on human evaluation preferences.
 - Maybe the head to head Chatbot Arena might work to mitigate that, but it can be bottlenecked by the number of people that vote.
 - Plus head to head battles won't pinpoint to the source of difference in scores.
- Who are the stakeholders in AI Evaluations?
 - **Productionisers:** people making models for a purpose that want to make sure it performs **robustly, safely, and responsibly**.
 - Hard to do automatically as it's a higher skill level focused. **Human evaluations** are used and are expensive.
 - For productionisers, it is **less relevant to score well on benchmarks that test specific skills in isolation**, or in a context different from the one the model will be deployed in
 - **Researchers:** the opposite of what I said for productionisers
-

Dimensions in Choosing Evals

- Generation vs Ranking
 - Multiple choice (ranking or classification tasks) or direct evaluation of generated text (free-form)
 - Ranking can be multiple choice, y/n, etc...
 - Can be automatic and cheap and useful in model dev to run continuously during iterations
 - Disadvantage is resemblance of scenario being modeled, but not the exact scenario; plus data contamination of the evaluations being part of the training data.
 - Direct open-ended evaluations are harder to do automatically... especially for non-deterministic answers. Human evaluation is the gold standard and suffer from poor replicability and disagreement around criteria and measurement and quality judgment from workers.
- Prompt Handling

- The variation of the prompt used can lead to the models giving out varied outcomes from the slightest arbitrary choices. Eg. choice of wording, number of prompts, and the way prompts are formulated and formatted. (A, B)
 - Maybe looking through the best case scenarios before computing evals to choose prompts that models respond well towards
 - Prompt engineering for evals depends on the stakeholders and; situations. For builders, re-engineering the prompt to choose the best performer can lead to seeing the upper-bound of the performance task and potentially re-routing similar prompts to the best performing one. However for performance researchers aiming to achieve AGI, prompt engineering scores are less informative as variations show how far off we are from corrective intelligence.
 - Also if a model responds unexpectedly to prompt-format changes in a relatively simple ranking test, what does that imply for the reliability of scores on tests for bias or toxicity;
- Handing Outputs
 - Especially to ranking ones. It's going to become increasingly easy to assign higher eval scores to models that are eval-instruction-tuned. So a simple Y/N might not be the move... Maybe looking at the assigned probabilities to each option (either in it's decision or by passing the prompt multiple times and seeing the probability of choice for each option) could be useful.
 -
- **Data Contamination**
 - Relationship between training and evaluation data.
 - If you're an AGI researcher, this is a no-go. Testing on stuff the model might have already seen counteracts the test. And not indicative of hyper-generalization and pattern matching.
 - There was a reported score of accuracy of GPT4 on HellaSwag evals that found 36% of reasoning questions were wrong, mislabeled or incomprehensible examples. [HellaSwag, or HellaBad? 36% of this popular LLM benchmark contains errors](#) that show the model arrived at the same incorrect outputs is no proof. So, maybe this is at least a strong indication that it may have arrived at those answers by memorizing its training set at least to some extent. Unless one is fairly certain that particular forms of generalization are not needed – i.e. the model will simply not encounter scenarios not covered directly by the training data frequently enough for it to matter – it may be wise to get an estimate of the extent to which high scores are driven by data contamination. It may seem unlikely a model has encountered a particular input before, but the internet is big, and searching it is hard.

- How much overlap between a training and evaluation example should there be for it to count as 'contaminated'? There are no easy solutions to these problems, approaches could include burning parts of common evaluation... However, to even get close to start addressing this insane issue, requires access to training data, they are in many cases not accessible to people outside of the ones that trained the models in question.
- Responsible LLMs
 - Even completely ignoring the potential consequences for humans and society, mistakes uttered by LLMs can have a direct impact on the viability of the product as well as the company itself
 - Despite the importance of this topic, broadly evaluating accuracy, fairness, toxicity, factuality and related topics remains incredibly difficult, both from a practical and philosophical perspective. The stakes of addressing it properly, however, are high, both for society and for model deployers directly: neglecting to properly evaluate these kinds of things can have a lot of inadvertent consequences for regulations, potentially making it impossible for the company to deploy similar models in their products.

Questions

- How far can automatic evals get us?
 - Human evals are expensive.
 - Are there ways to automatically evaluate higher-level skills?
- Data contamination
 - How should that be measured?
 - What is the impact of memorisation on benchmarks?
-

(1) Open LLM Leaderboard or HELM

Maintainer:

- Hugging Face & Stanford respectively

Summary:

- A sortable set of models, open and closed, evaluated on mostly academic benchmarks. Covers core scenarios such as Q&A, MMLU (Massive Multitask Language Understanding), MATH, GSM8K (Grade School Math), LegalBench, MedQA, WMT 2014 along with other benchmarks.
- Largely open source evals that have been developed over the years by various researchers.
- It encourages leaderboard hacking and faster saturation of academic benchmarks and wasn't well thought through.

Developer goal:

- General AI developers looking for models to build on will come here first, understand per a given size, which models perform the best and then leverage the result for their research, an application, fine tune for a custom application.
- One thing that's clear however is that there is a stage of disillusionment that the community is in when it comes to ****actual**** performance of models and how each rank on academic evals. This has driven the development of other leaderboards such as Chatbot Arena (see later).

(2) Hallucinations Leaderboard

Maintainer:

- Hugging Face

Summary:

- It evaluates the propensity for hallucination in LLMs across a diverse array of tasks, including Closed-book Open-domain QA, Summarization, Reading Comprehension, Instruction Following, Fact-Checking, Hallucination Detection, and Self-Consistency.
- The evaluation encompasses a wide range of datasets offering a comprehensive assessment of each model's performance in generating accurate and contextually relevant content.

Developer goal:

- Have a starting point to understand which models hallucinate the least.
- There is some overlap with other available leaderboards here and, given without RAG or search integrated, these models will ****always**** hallucinate, this leaderboard is less interesting for developers.

(3) Chatbot Arena / LMSys

Maintainer:

- Together AI, UC-Berkeley, Stanford and several others.

Summary:

- A benchmark platform for large language models (LLMs) that features anonymous, randomized battles in a crowdsourced manner.
- Some, like Karpathy, find it to be the best test of a model's performance as the arena is much more dynamic and **compares models head to head**.
- Reliable in telling you which model is better, but can't get into the details of specific capabilities (eg. gemini vs gpt4, where gemini had internet access and caused confusion)

Developer goal:

- Understand how models will perform in a **real** environment that challenges in ways that status datasets / prompts can't.
- Like #1 above, this may influence the developer's starting point for further work.

(4) The MTEB leaderboard

Maintainer:

- Hugging Face & Cohere

Summary:

- MTEB or "Massive Text Embedding Benchmark" contains 8 embedding tasks covering a total of 58 datasets and 112 languages. Through the benchmarking of 33 models on MTEB, establish the most comprehensive benchmark of text embeddings to date.

Developer goal:

- Understand which models could perform best in applications where retrieval augmented generation (**RAG**) is used.

(5) Artificial Analysis

Maintainer:

- Independent / Startup

Summary:

- Artificial Analysis provides benchmarks & information to support developers, customers, researchers, and other users of AI models to make informed decisions in choosing: **Which AI model to use for a given task, and Which hosting provider to use to access the model.**

Developer goal:

- Understand cost, quality and speed tradeoffs across models and model providers (e.g. OpenAI, Microsoft Azure, Together.ai, Mistral, Google, Anthropic, Amazon Bedrock, Perplexity, Fireworks, Lepton, and Deepinfra.)

(6) Martian's Provider Leaderboard

Maintainer:

- Martian (startup)

Summary:

- Martian's provider leaderboard collects metrics daily and tracks them over time to evaluate the performance of LLM inference providers on common LLMs.

Developer goal:

- Understand **quickly cost vs. rate limits vs. throughput** for a variety of LLM providers and optimize for his or her application.

(7) Enterprise Scenarios leaderboard

Maintainer:

- Patronus, Hugging Face

Summary:

- Enterprise Scenarios leaderboard evaluates the performance of language models on real-world enterprise use cases (at present 6 benchmarks - Finance, Legal, etc...)
- FinanceBench (accuracy not retrieval), Legal Confidentiality (reason over legal cases), Writing Prompts (engagingness), Customer support Dialogue (relevance), Toxic Prompts (generation), Enterprise PIII (business sensitive info generation)

Developer goal:

- This leaderboard is still very much **nascent** as you can see from the **lack of public models competing**.
- Overall though this set of evals would provide developers with an idea of which base models to use as a starting point for a given task. Again very very early.

8. ToolBench

Maintainer:

- SambaNova Systems

Summary:

- A tool manipulation benchmark consisting of diverse software tools for real-world tasks.

Developer goal:

- Illustrate which available models perform best for action generation as described in natural language. Note it seems like this project is also a great resource for broader developers for how to implement a particular tool usage (i.e. plug-in) for their model.

SANDBOX FOR DIFFERNET EVALUATERS AND SEE WHICH PERFORMANCE WHERE...

- SORT OF META EVALAUTOR BY EVALAUTING THE EVALATUAION EVALATIONERO RS

<https://artificialanalysis.ai/>

Unrelated to patronus, but mainly hebbia:

- knowing the prompt sensitivity of models could be valuable information for model deployers.
- Furthermore, differences in the way that prompts are selected make it impossible to compare performance for different models: is model A better than model B because it is better at the task, or did the developers of model A simply consider more prompts in their search?

- Whatever prompting scenario is used, it is extremely important that prompt-selection efforts as well as prompts themselves are well-documented. Not doing so is not only harmful for the reproducibility -- in the process wasting the time of countless scientists that need to back-engineer the prompts with which reported scores were obtained -- but also limits researchers to better understand what models are capable of.

Evaluating outcomes

- Similar to prompt engineering, using extracted, calibrated probabilities rather than generated text may be more informative about the upper-bound of what models can do than about what a model would actually do in practice, and it is interesting to consider the implications of that for the different stakeholders. For reasons analogous to the ones presented for prompt engineering, using calibrated log-probs may be an apt evaluation method for performance researchers: it gives a quick indication of what kind of knowledge the model actually contains and therefore constitutes an indicator of progress. For productionisers, however, it may be more appropriate to focus on generated output. While several preprocessing options may be conceived to address the variability that occurs as a consequence of prompting, the fact that a model has the correct answer somewhere in its probabilities is not particularly useful for a model deployer if there is no way to extract that from the model without knowing the answer in advance. As such, comparing the probabilities of answer A, B and C instead of checking if a model actually generates A, B or C may further distance the evaluation scenario and the scenario in which a model will be deployed. For an AGI researcher, using log-prob evaluations will evidently not suffice either: if a model in a multiple choice task assigns a higher probability to option A than to option B, but it generates the letter D, an AGI researcher would likely argue that the model did not even really understand the task it had to do. Similarly, in many cases, normalising probabilities may counteract the very idea by which in particular NLI tasks were constructed: predicting the relationship between two sentences independently of how likely the sentences themselves are is a crucial part of such tasks. In sum, focusing evaluation directly on generated model output might thus be a more suitable option, even if that results in lower scores.

Also from the gemini 1.5 paper the request for multimodal evaluations that perform well for Long Context Evaluation was interesting.

- Patronous Current evals fail with: Existing evaluations are increasingly strained by the new and rapidly advancing capabilities of large multimodal models. They typically focus on individual modalities and/or are restricted to tasks with shorter context lengths. Hence, there is a growing need for benchmarks which exemplify the nuanced requirements of real world long mixed-modality use cases.
 - Qualitative long-context multimodal evaluations: manually probe and stress-test the model's long-context abilities, especially for novel capabilities where no quantitative benchmarks exist

- Especially relevant with their long-context evaluation call to action: Thus, given the limitations of existing benchmarks and the challenges of human annotation, there is a pressing need for innovative evaluation methodologies. These methodologies should be able to effectively assess model performance on very long-context tasks while minimizing the burden of human labeling. To begin addressing some of these concerns, we recommend researchers and **practitioners adopt a “multiple needles-in-haystack” setup for diagnostic evaluations which we observed to be signal-rich and more challenging compared to its “single needle” counterpart.** We believe there is room for new benchmark tasks based on new or improved automatic metrics that require complex reasoning over long inputs (both human and model generated). This is also an intriguing research direction for creating challenging evaluations that stress more than just the retrieval capabilities of long-context models. We will be continuing the development of such benchmarks for realistic and comprehensive evaluation of model capabilities in the multimodal space. By addressing these open challenges and developing new benchmarks and evaluation methodologies, we can drive progress in the field of very long-context AI models and unlock their full potential.

▼ Anomaly Detection in Timeseries Data

[Database View \(Status and Dates Written\)](#)

Tech
Finance
Arts

Description

For general use of for internal needs for a prospecting tool monitoring time-series trends, identifying inflections/anomalies, and present filters. I imagine it's a tool that lets you put in a metric or query and it generates a basic time-series prediction and then can notify you if it's out of bounds given input bounds and alert thresholds. I was playing around with the idea last night (attached) and hosting it on streamlit for easy interactions.

[Timeseries_Anomaly_Detection_to_Streamlit \(1\).ipynb](#)

PS: (1) used data like DAU, Funding Rounds, Acquisition Amounts, NYC taxi riders etc as I thought they're similar/relevant to the type of data most of the readers here ingest...

IPYNB Files

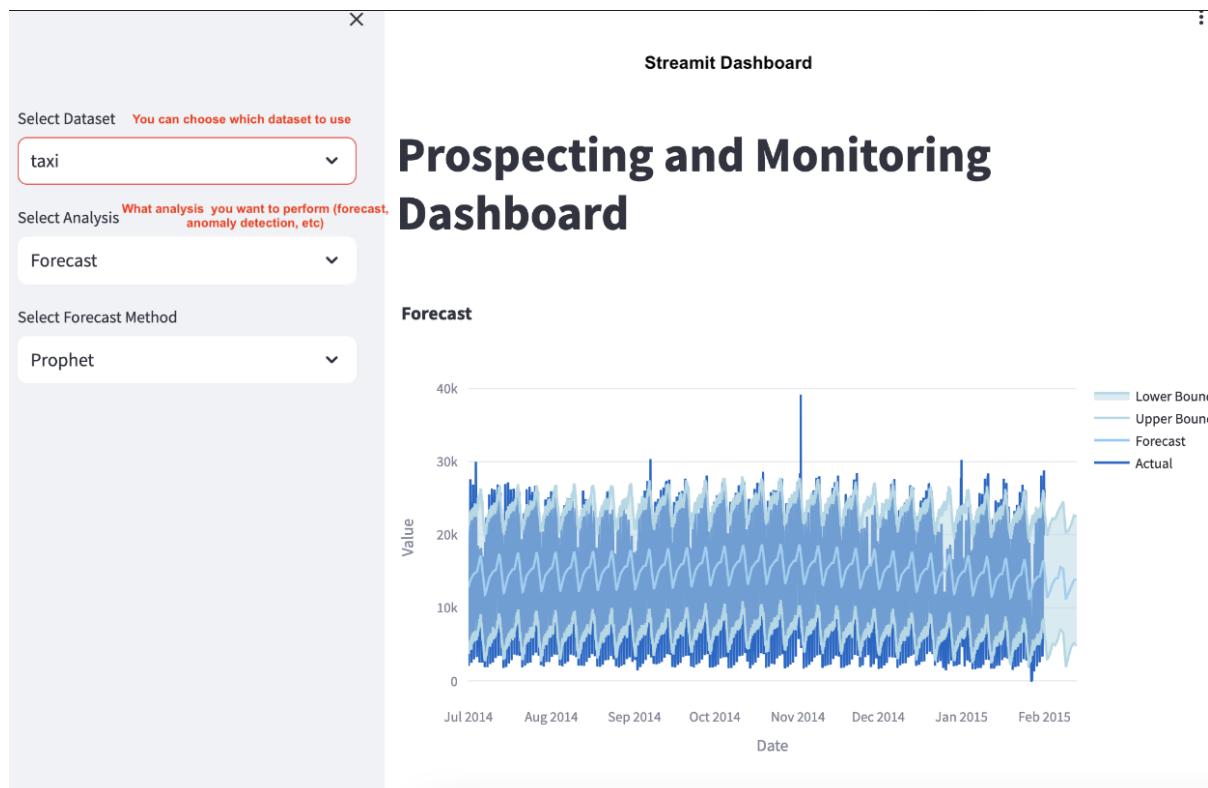
[Detection through ADTK + Isolation_Forest \(4\).ipynb](#)

[Detection_through_Darts.ipynb](#)

[Detection_through_lagllama.ipynb](#)

[Detection_through_Transformers.ipynb](#)

Mockup



Gists for Visibility

<https://gist.github.com/amenti4k/43da3e70407c7933ca2833667455bb18>

<https://gist.github.com/amenti4k/988d73dc8a0dd4fc535427b789827cbe>

<https://gist.github.com/amenti4k/dc286ae3dd187f4414a2c4c99b6deac0>

<https://gist.github.com/amenti4k/9a251ca1f1eed1dc79eb1b698175d97f>

<https://gist.github.com/amenti4k/63f48863ebc49246b938589e1e8f37c4>

▼ Beyond Prompts

- All I want to do is steering towards acceptable results rather than just tweaking prompts and hoping for the best — a judicious **balance of constraints and freedom**.
- Finding interaction patterns that give more calibrated control could be key. How can we discover interfaces that unlock deeper and more tailored integrations between users and generative **models beyond sentence prompts?** This could significantly augment creative and knowledge work.

The Curse of Indirection

- Current interfaces for working with generative AI models are **indirect**—we manipulate models mainly through text prompts and conversations. This adds friction and distance between the user's intent and the model's output.
- Current text prompts place generative models at arm's length, like trying to steer a car from the passenger seat. **More integrated, direct ways of manipulating models could improve workflows.** More integrated interactions could provide proper driver's seats for precise guidance.
- Quoting Kate Compton's Casual Creators theory "the possibility space should be **narrow enough to exclude broken artifacts... but broad enough to contain surprising artifacts as well**. The surprising quality of the artifacts motivates the user to explore the possibility space in search of new discoveries, a motivation which disappears if the space is too uniform"
- Context menus inside documents that let users branch out of their current vertical by highlighting texts/keywords could be way to overcome indirection.
 - ▼ Example of process that might generate better value: **Hyperlinks on results. Clicking** helping to expand the topic, based on the prompt being discussed, then clicking back

minimizing it. Word exploration through clickable words that function as ever expanding tree toggles

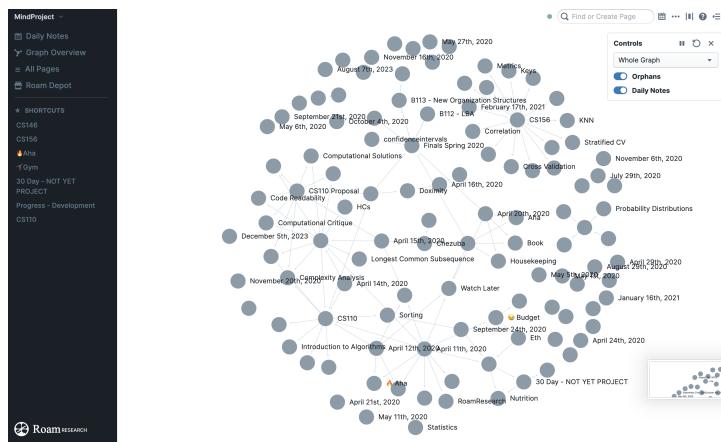
right?!

- Even within the space of text based interaction, we want to keep the **lineage of information** that changed over time instead of overwriting the fact. Forexample using Rich Hickey's perspective on information updating theory, "If my favorite color was red and now it's blue, we don't go back and change the fact that my favorite color was red to be blue – that's wrong. Instead, we **add a new, updated fact** that my favorite color is now blue, but the old fact remains historically true."

Exploring Latent Spaces

- Moving through latent spaces quickly and viscerally is an alternative to conversational prompts. This ties to the idea from that **we lack tactile, direct ways to guide text generation.** New interaction techniques that let users directly **manipulate directions** and vectors in latent space could unlock more creative possibilities. Traversing latent idea spaces via prompts resembles blind navigation through textual adventure games. New interaction paradigms could make exploration more immersive, like being in the drivers seat of the tools using these vectors to explore/discover in latent space.
 - What most traditional apps that sit on top of large amounts of data do is, how do you take a commodity in a database and layer curation and recommendation in ways that are more usable and friendly than just giving people a search box and pushing them out of the door? Is there a way to add **horizontal expansion** to search instead of vertical digging that requires reformatting inserts? How do you break out of hierarchical directories that don't scale (ie. Yahoo's directory) — even when the hierarchies are just ranked search results from the users' prompts?
- I keep referring back to the covid times where I started using Roam Research for my note-taking. I was in college and had time. Back-propagation by directly playing around with interfaces. I didn't start out being a programmer so I've always wondered about **how to intuitively control end products to change the source code.** Further extending this with what I said about language, how can we use the newfound abilities of coding on command to back-propagate information processing? So like instead of going on my weather app and searching through when it's warm enough to leave my apt in cold nyc Dec without a jacket, moving the temperature slider to a higher degree to back-propagate the dates.

I keep thinking of what the nested knowledge graphs of `roam.research` look like if they were autogenerated instead of us manually generating interlinks. Learning would be awesome!



- This is my roam pages networked graphics during covid when I had the time to interlink notes. It was fun and useful, but never ended up working for me due to the intensive writing process whenever typing to include “[[]]” whenever trying to interlink topics and having to manually remember what to even link.
- Especially useful when the tool picks up on **notes that are proper names** that need to be clarified further...
 - It can help me notice connections between ideas in my notes that I wouldn’t have even thought to make myself, even if I were trying to find interesting notes to link together. With a smarter system, a similar interface could even automatically discover and show links from your notes to high-quality articles or online sources that you may not have seen yet, automatically crawling the web on your behalf.

Going back to the analogy of driving cars, in addition to giving you the seat and a steering wheel, it’s allowing you to have a windshield to look across and see where you want to maneuver!!

Balancing Guardrails and Possibilities

- As noted in the initial thoughts, providing guardrails for safety while preserving expansive capability spaces is an important challenge. At the end of the day, we need windshields for a reason! Permitting **expansive possibility spaces** risks accidents or misuse. Even the drifting of attention. However, back-propagating user edits to tune outputs that could strike this balance.

Anyways, tying back this to current professional parsing tools I think **finance, legal, and medical sectors** deal with highly complex and structured data sets. Outside of the commonly thought of reasons on working with these data structures (like the high stakes decisions, regulatory compliance and precision needed, and room for automation/personalization), I think **the complexity offers fertile ground for experimenting with innovative interaction models** to manage, interpret, and manipulate such data effectively given their pre-given formatting.

Let me know if you would like me to elaborate or focus on any part of this synthesis further! I’ll leave with this: **Static information media severely limits what ideas we can express** and

manipulate. We're limited by how much we can conveniently represent, and so much thinking is still trapped inside the head. Dynamic, interactive media could empower entirely new realms of thinking.

Anecdotally, one summer at fb messenger my main role was on message search and ability to surface it well. I look at what is the best way to give the prompter something they want — **even before they realize they want it.** It all started with looking at simple descriptive stats about usage for in chat searches. People commonly searched for numbers, emails, passwords, or dates/locations. What if there's a way to use people's current usage flow, to add layers that guide to more discovery instead of just waiting for the user guided flow.

Obviously the worry here is not to **overwhelm the user** by providing buttons/flows they didn't ask for. But I believe there is a world where it can empathetically be done!

Another worry might be tools like Harvey or Hebbia using users VDR to bring up knowledge graphs and predefined prompts that might **seem intrusive** and not-secure. I hope the only things standing between our current state and when this becomes the norm is some time and better enterprise ai security systems.

This is just wonderings I've had just written down to help me visualize my thoughts. Regardless, let me know what you think or lines/topics you'd want to explore further.

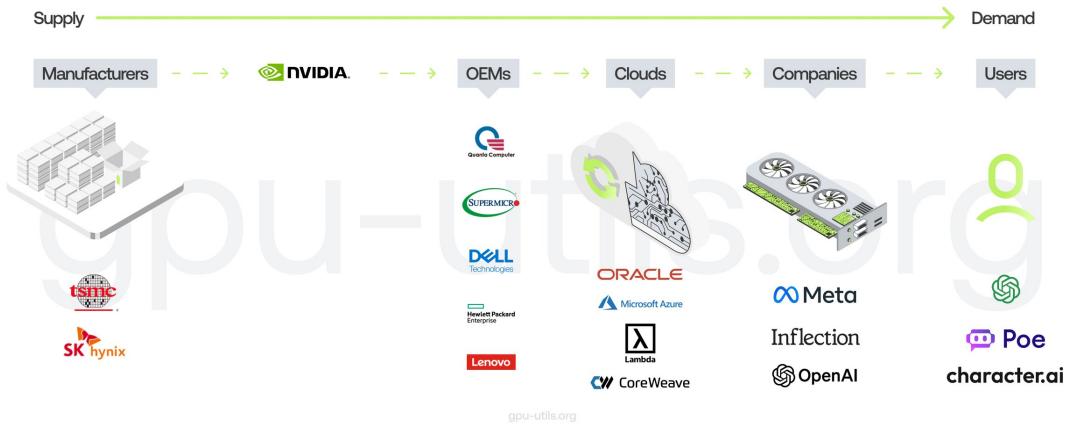
▼ wtf are gpus?

A [long post exploring the supply and demand of GPUs](#) came out last week, and it is an excellent snapshot of the Nvidia supply bottleneck. If you have time, you should read the entire article, but if you are less interested in the more technical aspects, and want to understand both the demand forces and supply constraints, I tried capturing all its main points below.

(All the credit goes to Clay Pascal and the GPU Utils team for their research, analysis, and excellent article).

If you'd like to read community reactions to the article, I also recommend this [Hacker News](#) thread.

The Journey of GPU Supply and Demand ([via GPU Utils](#))



Why is this topic so important?

GPUs are the hardware that makes LLMs possible. It's what powers the software applications that we all see and use. Without the computing power that these chips afford us, we wouldn't be able to train and run most, if not all, AI applications that we're seeing today.

Right now, the supply of these GPUs are running out. This has many consequences: (1) it slows down how fast we can continue building, and (2) access becomes limited to the privileged few who effectively become gatekeepers to AI development.

It's very possible that we have a feedback loop right now, where scarcity causes GPU capacity to be perceived as a moat, which then causes more GPU-hoarding, which in turn exacerbates scarcity. This is dangerous and it makes innovation in the space increasingly more expensive, and thus, concentrated behind the FAANGs.

Levels of excitement and momentum in the startup ecosystem are at an all time high, and justifiably so. However, I think everyone, even the ones just tinkering with applications on top of this [new LLM stack](#), should be aware of the real-world bottleneck limiting speed and cost of innovation in the space. Otherwise, too much energy may go to waste by having a limited perspective or staying focused on the wrong things.

Tracing the Journey of GPU Supply And Demand through the lens of ChatGPT, the LLM product that's found the strongest product-market fit:

1. Users love ChatGPT. It's probably making \$500M+ in annual recurring revenue.
2. ChatGPT runs on the GPT-4 and GPT-3.5 APIs.
3. The GPT-4 and GPT-3.5 APIs need GPUs to run. Lots of them. And OpenAI wants to release more features for ChatGPT and their APIs, but they can't, because they don't have access to enough GPUs.
4. They buy lots of Nvidia GPUs through Microsoft/Azure. Specifically the GPU they want most is the Nvidia H100 GPU (their latest, more powerful GPU).

5. To make H100 SXM GPUs, Nvidia uses TSMC for fabrication and uses TSMC's CoWoS (Chip on Wafer on Substrate) packaging tech and uses HBM3 primarily from SK Hynix (high bandwidth memory supplier).

OpenAI isn't the only company that wants GPUs (but they are the company with the strongest product-market-fit that wants GPUs). Other companies are also wanting to train large AI models. Some of these use cases will make sense, but some are more hype driven and unlikely to get product-market-fit. This is pushing up demand. Also, some companies are concerned about not being able to access GPUs in the future so they're placing their orders now even when they don't need them yet. So there's a bit of "expectations of supply shortages create even more supply shortages" going on.

The other major contributor to GPU demand is from companies that want to create new LLMs. Here's the story of GPU demand with respect to companies wanting to build new LLMs:

1. A company executive or founder knows that there are big opportunities in the AI space. Maybe they're an enterprise that wants to train an LLM on their own data and use it externally or sell access, or maybe they're a startup that wants to build an LLM and sell access.
2. They know they need GPUs to train large models.
3. They talk with some set of people from the big clouds (Azure, Google Cloud, AWS) to try and get many H100s.
4. They find out that they can't get a big allocation from the big clouds, and that some of the big clouds don't have good networking setups. So they go and talk with other providers like CoreWeave, Oracle, Lambda, FluidStack. If they want to buy the GPUs themselves and own them, maybe they also talk with OEMs and Nvidia.
5. Eventually, they acquire a lot of GPUs.
6. Now, they try to get product-market-fit.
7. In case it's not obvious, this pathway isn't as good - remember that OpenAI got product-market-fit on much smaller models and then scaled them up. But, now to get product-market-fit you have to be better than OpenAI's models for your users' use-cases, so to start you will need more GPUs than OpenAI started with.

Access to GPUs has become so pressing that a new market is forming aimed at helping developers find available GPUs across all providers, with startups such as Brev Dev. Other companies in this space, such as Exafunction, are also doing GPU orchestration, i.e. optimizing efficiency and cost of running these models.

Demand Side

Interesting facts, insights, or open questions:

- In short: Yes, there's a supply shortage of H100 GPUs. For companies seeking 100s or 1000s of H100s, Azure and GCP are effectively out of capacity, and AWS is close to being out. This "out of capacity" is based on the allocations that Nvidia gave them.

- Who Needs/Has 1,000+ H100 Or A100 (previous gen)?
 - Startups training LLMs
 - OpenAI (through Azure), Anthropic, Inflection (through Azure and CoreWeave), Mistral AI
 - CSPs (Cloud Service Providers)
 - The big 3: Azure, Google Cloud Platform (GCP), AWS
 - The other public cloud: Oracle
 - Larger private clouds like CoreWeave, Lambda
 - Other large companies
 - Tesla, Jane Street, ByteDance, Tencent, etc.
- Which GPUs Do People Need?
 - Mostly H100s. Why? It's the fastest both for inference and training for LLMs. (The H100 is often also the best price-performance ratio for inference, too). Specifically: 8-GPU HGX H100 SXM servers.
- How Much Do These GPUs Cost?
 - 1x DGX H100 (SXM) with 8x H100 GPUs is \$460k including the required support. \$100k of the \$460k is required support.
- How Many GPUs Are Needed?
 - GPT-4 was likely trained on somewhere between 10,000 to 25,000 A100s.
 - Meta has about 21,000 A100s, Tesla has about 7,000 A100s, and Stability AI has about 5,000 A100s.
 - Falcon-40B was trained on 384 A100s.
 - Inflection used 3,500 H100s for their GPT-3.5 equivalent model (they'll have 22k operational by December).
 - GCP has approx 25k H100s. Azure probably has 10k-40k H100s. Should be similar for Oracle. Most of Azure's capacity is going to OpenAI. CoreWeave is in the ballpark of 35k-40k H100s - not live, but based on bookings.
- How Many H100s Are Most Startups Ordering?
 - For LLMs: For fine tuning, dozens or low hundreds. For training, thousands.

What's stopping companies from using AMD GPUs? Why aren't there other AI chips companies?

This question has received George Hotz's latest attention. Aside: George is known as the guy who hacked the first iPhone, and more recently known for being the last Twitter intern.

George just announced that he started a new company called [Tiny Corp](#). His claim: the only way to start an AI chip company is to start with the software. The current crop of AI chip companies [failed](#). Many of them managed to tape out chips, some of those chips even worked. But not a single one wrote a decent framework to use those chips. They had similar performance/cost to Nvidia, and way worse software. Of course they failed. Everyone just bought stuff from Nvidia.

While AMD stayed focused on the gaming industry, Jensen Huang (Nvidia's founder/CEO) took a bet on the up-and-coming proliferation of AI and [focused most of the company's resources](#) on becoming the leader in the space. Nvidia built CUDA, their software stack that allows programmers to interact with the hardware (GPUs).

It's not surprising that Nvidia has a head start and AMD's stack is immature. However, AMD is accelerating to catch up, using the heat in this space in their favor. We're starting to see some positive signals; MosaicML is exploring how to build LLMs with AMD instead of Nvidia: [Training LLMs with AMD MI250 GPUs and MosaicML](#).

How many H100s might companies be wanting?

OpenAI might want 50k. Inflection wants 22k. Meta maybe 25k (article says actually Meta wants 100k or more). Big clouds might want 30k each (Azure, Google Cloud, AWS, plus Oracle). Lambda and CoreWeave and the other private clouds might want 100k total. Anthropic, Helsing, Mistral, Character, might want 10k each. Total ballparks and guessing, and some of that is double counting both the cloud and the end customer who will rent from the cloud. But that gets to about 432k H100s. At approx \$35k a piece, that's about \$15b worth of GPUs. That also excludes Chinese companies like ByteDance (TikTok), Baidu, and Tencent who will want a lot of H800s.

There are also financial companies each doing deployments starting with hundreds of A100s or H100s and going to thousands of A/H100s: names like [Jane Street](#), JP Morgan, [Two Sigma](#), [Citadel](#).

With a limited supply, Nvidia could purely raise prices to find a clearing price, and are doing that to some extent. But it's important to know that ultimately H100 allocation comes down to who Nvidia prefers to give that allocation to.

I think it's possible we have a self-reinforcing cycle right now where scarcity causes GPU capacity to be perceived as a moat, which causes more GPU-hoarding, which exacerbates scarcity.

– Exec from a private cloud company

Supply Side

Interesting facts, insights, or open questions:

- Can Nvidia use other chip fabricants outside of TSMC for H100 production?

- Not really, at least not yet. They've worked with Samsung in the past. But on the H100s they only use TSMC. Implication is that Samsung can't yet meet their needs for cutting edge GPUs. They might work with Intel in the future, and Samsung again on cutting edge, but neither of those will be happening in the short term in a way that'd help the H100 supply crunch.
- How Long Does Production Take (Production, Packaging, Testing)?
 - Estimates: 6 months from production on a H100 starting to that H100 being ready to be sold to a customer.
- Who makes the memory on the H100s?
 - For the HBM3 memory itself, Nvidia uses either all or mostly SK Hynix. Not sure if Nvidia uses any from Samsung for the H100s and the article believes it's nothing from Micron for the H100s.
- What other pieces are used when making GPUs that may be bottlenecked?
 - Metal Elements: These are essential in the production of GPUs. They include:
 - Copper: Used in the creation of electrical connections due to its high conductivity.
 - Tantalum: Often used in capacitors due to its ability to hold a high electrical charge.
 - Gold: Used in high-quality plating and connectors due to its resistance to corrosion.
 - Aluminum: Frequently used in the heatsink to help dissipate heat.
 - Nickel: Often used in the coating of connectors for its corrosion resistance.
 - Tin: Used in soldering components together.
 - Indium: Used in thermal interface materials for its good thermal conductivity.
 - Palladium: Used in certain types of capacitors and semiconductor devices.
 - Silicon (Metalloid): This is the primary material used in the creation of semiconductor devices.
 - Printed Circuit Board (PCB): This is the board on which all the components of the GPU are mounted. It provides the electrical connections between the components.
- Who sells H100s?
 - Original equipment manufacturers (OEMs) like Dell, Hewlett Packard (HPE), Lenovo, Supermicro and Quanta sell H100s. And when you need InfiniBand, you'll need to speak directly to Mellanox (Nvidia's subsidiary since 2019).
 - So GPU clouds like CoreWeave and Lambda buy from OEMs and then rent to startups.
 - Hyperscalers (Azure, GCP, AWS, Oracle) work more directly with Nvidia but they are generally also working with the OEMs as well.

- And even for DGX (Nvidia's all-in-one solution for enterprise AI) you'll still buy through an OEM. You can talk to Nvidia, but you'll buy through an OEM. You won't do a purchase order directly to Nvidia.
- When Do Startups Build Their Own Datacenter Vs Doing Colocation?
 - For building a datacenter, the considerations are the time to build the datacenter, whether you have the people and experience in hardware, and that is capex expensive (major, long-term costs).
- How do the big clouds compare?
 - The sentiment is that Oracle infrastructure is less reliable than the big 3 clouds. In exchange, Oracle gives more tech support help and time
 - The main big differences at the large clouds are:
 - Networking (AWS and Google Cloud have been slower to adopt InfiniBand because they have their own approaches, though most startups looking for large A100/H100 clusters are seeking InfiniBand)
 - Availability (Azure's H100s are mostly going to OpenAI. GCP is struggling to get H100s.)
- Nvidia seems to tend to give better allocations to clouds that aren't building competing machine learning chips. (This is all speculation, not hard facts.) All of the big 3 clouds are working on machine learning chips, but the Nvidia-alternative offerings from AWS and Google are already available and taking dollars that might've gone to Nvidia.
- Oracle was the first to launch A100s, and they worked with Nvidia to host an NVIDIA-based cluster. Nvidia is also a customer of Azure.
- Which companies use which clouds?
 - OpenAI: Azure.
 - Inflection: Azure and CoreWeave.
 - Anthropic: AWS and Google Cloud.
 - Cohere: AWS and Google Cloud.
 - Hugging Face: AWS.
 - Stability AI: CoreWeave and AWS.
 - Character.ai: Google Cloud.
 - X.ai: Oracle.
 - Nvidia: Azure.

How can a company or cloud service provider get more GPUs?

The ultimate bottleneck is getting allocation from Nvidia. They have an allocation that they give per customer. But for example, Azure saying, "Hey we would like 10,000 H100s for use by Inflection" is very different from Azure saying "Hey we would like 10,000 H100s for Azure's cloud." That's to say that Nvidia cares about who the end-customer is, and so cloud providers might be able to get an extra allocation for a specific end-customer if Nvidia is inclined. Nvidia also wants to know who that end-customer is, as much as possible. And they prefer customers with nice brand names or startups with strong pedigrees.

Nvidia would prefer not to give large allocations to companies that are attempting to compete directly with them (AWS Inferentia and Tranium, Google TPUs, Azure Project Athena).

But ultimately, if you put the purchase order and money in front of Nvidia (e.g. committing to a bigger deal and more money up front and show that you have a low risk profile) then you'll get a larger allocation than others get.

It's likely the shortage of GPUs will persist through the rest of 2023, but it's difficult to predict what happens next. If you are interested in getting more information like this, you should [sign up to get notified](#) about LLM Util's new posts.

▼ Suddenly Popular LLMOps

LLMops

Sometimes, all of sudden, micro-markets emerge. They can be triggered by all sorts of things, for example an external change (COVID) or a new technical capability (LLMs). The current LLMOps/PromptOps space is an instructive example. Over the last year, the number of developers experimenting with AI model APIs has 1000X'd.

The cycle to date has been something like this:

Models at scale have emergent behaviors that are magical and shocking.

Consumers experienced DALLE2, ChatGPT, and a small number of LLM products gained real traction rapidly (Copilot, Jasper, Midjourney, Character).

Startups have flocked to leverage these capabilities, VCs are funding them like it's 2021. Many incumbent technology leadership teams are excitedly, anxiously resourcing AI projects. These developers all start by tinkering: they try different prompts, chain together model API calls, connect to other non-AI services, and integrate with input data sources. OSS frameworks such as LangChain and LlamaIndex, and a significant cohort of YC companies have already emerged to solve some piece of this problem. A million developers are trying to do the same thing, experiment and ship a prototype. Entrepreneurial developers see an opportunity. The billion dollar question is whether all this interest leads to any durable market.

The history of software features many legendary companies that started with an elephant of a vision almost too big to take the first bite of (Figma, who collapsed several categories of software and put them into the collaborative web to solve end-to-end for product designers). But it is also populated by companies that iterated to platforms, starting with a timely wedge

(Hubspot, which expanded from SMB content marketing to the only real contender to Salesforce).

We believe great companies can emerge from the morass of spaces like "LLMOps." But those that do will be teams that see the wedge for what it is, rather than misreading immediate momentum and interest for durable value. The distance between Github stars and Twitter likes and at-scale deployments and six figure enterprise contracts is very far. Solving an easy but acute problem in a temporary market, faster than others do, can be a smart entry point to get momentum. All things are possible for a startup with momentum, money, and the right management team.

When everyone sees the same needs, the bar for understanding those needs and executing on them goes up. The question is not, "Do developers want LLMOps?" but instead, "Which segment of those users do I focus on? What do they really need, and in what order? What will make the product easy to adopt, and what objections will I face? What architecture will support those users, and what compounding advantages can I build?"

AI is a landscape of shifting sands. What developers want today is not what they'll want in six months, and what they need to build demos is not what they need in production, is not what they'll need for integration into existing products. But demos could be the path to distribution. Marching in lockstep with customers along the path to market maturity requires being even more "niche" in an already small market because there are segments even now. The closer you are aligned with where some set of customers are today, the more customer trust you can build, the more likely you are to find demand others don't understand, the better chance you have of building a very important company.

At the beginning of a market, no one really knows what user needs are. Founders who have solved a problem themselves, ahead of the crowd (or a previous iteration of it) have some advantage. But because the market is evolving, founders who are learning from customers, who launch and then have the resolution of conversation necessary to really develop a product, have even more advantage. I've often been surprised how common it is for startups to have an insufficient depth of understanding of customer problems, or to misread the signals from customers. Especially when working with friends and early adopters, people are inclined to be nice. If a smart and charismatic team describes a high-level problem they face reasonably accurately, they'll nod assent, nicely. "Would you like to lose weight?" is a very different question than, "will you lose weight by eating $\frac{1}{3}$ fewer calories, not drinking socially, and prioritizing workouts four days a week?" Customers want to solve problems. They may not picture the roadblocks to adoption and tradeoffs. They may not be willing to be directly skeptical. Here is where increasing resolution of conversations, forced prioritization and asking for the sale all provide better signal.

Sometimes, emboldened by the strength of immediate need, and feeling the pressure to raise money and execute quickly in a noisy market, founders will be quickly drawn to "defensible technical depth" as their narrative to investors. The risk is that they're not yet sure it's true, but they say it enough to convince themselves of a world model that's wrong. Counterintuitively, recognizing that no part of solving the immediate problem is hard forces a more useful ongoing

search and paranoia. Defensibility is overstated for most early-stage startups. It is wrongfully sought by investors, too.

The problem with the "sell picks and shovels during a gold rush" analogy is that picks and shovels are fungible, and software products are not. Eventually, defaults emerge. The risk of solving easy problems is that they're easy for other people to solve too. They can be solved by incumbents with a distribution advantage, or by other startups.

Leadership even in "temporary markets" is a valuable position, and "easy problems" can still be good entry points for startups to leverage. Almost any growing problem ends up deeper than it first appears.

▼ Vector Embeddings - Hype from Excess Dry Power?

Embeddings – A Hype Cycle Fueled by Excess Dry Powder?

At the bottom of everything that I'm trying to do here, what I'm trying to do is evaluate whether AI companies are worth leaving everything being and betting on?

Reflexivity Framework

I want to start off with the idea of reflexivity as I assume the best investors put their earnings and future (skin in the game) by predicting how the future goes. This relates as I'm looking if there is anything material in the AI space that will change the career direction I take. Investors don't base their decisions on reality, but rather on their perceptions of reality instead. A framework is essential when looking at new technologies and the ecosystem it creates.

However, their actions from these perceptions have an impact on reality, or fundamentals, which then affects investors' perceptions and thus prices. The process is self-reinforcing and tends toward disequilibrium, causing prices to become increasingly detached from reality – ie. crypto.

People get used to things. People think about the world through the lenses provided by the status quo of the things they use. Then when the world changes, sometimes whole new ideas are possible. The strongest example of this is probably the Web. It enabled all kinds of ideas that people didn't think of before. The network of interconnected computers provided a new mental model for them to work from to invent new things.

Social media didn't immediately come with the web. Why not? It takes time for the new reflexive part of an innovation to arrive. To understand what is fully possible under the new technology paradigm, some people need to have **worked in it natively** for a few years so that they begin to break down the status quo way of thinking.

Defensibility in Building

From a technical perspective it's a huge breakthrough that will have lasting impacts. As technology makes doing more stuff faster and easier, it's increasingly difficult to find areas of long-term defensibility in business models. The key position investors seem to be taking is that "context layers" that take these generative tools and put them into some point solution of a workflow is the place to make a bet.

It's hard to make these defensible for two reasons:

First, there will be too many players because the barriers to entry are low and that drives a competitive dynamic that is unfavorable to investors.

Second, they risk competition with the foundation models themselves as those models improve. Not only could OpenAI boot your company off of their API, but they could also improve upon their model faster than you can build out the middle layer - rendering your improvements useless in a matter of days with a massive new update.

Sometimes, emboldened by the strength of immediate need, and feeling the pressure to raise money and execute quickly in a noisy market, founders will be quickly drawn to "**defensible technical depth**" as their narrative to investors. The risk is that they're **not yet sure it's true, but they say it enough** to convince themselves of a world model that's wrong.

Counterintuitively, recognizing that no part of solving the immediate problem is hard forces a more useful ongoing search and paranoia. Maybe defensibility is overstated for most early-stage startups. At seed stage, the only defensibility is the quality of founders. **Also, It's actually irresponsible to *not* leverage GPT – similar to mobile/cloud.** *Startups are often a spread trade on new innovations before wider adoption. And especially salient with such a general purpose technology like LLMs — a rising tide.*

Most \$10B+ companies seem defensible now, but it took them several years...execution is the only real moat. It is wrongfully sought by investors, too. The problem with the "sell picks and shovels during a gold rush" analogy is that **picks and shovels are fungible, and software products are not. Eventually, defaults do emerge.** The risk of solving easy problems is that they're easy for other people to solve too. They can be solved by **incumbents with a distribution advantage**, or by other startups. That's why leadership even in "temporary markets" is a valuable position, and "easy problems" can still be good entry points for startups to leverage. Almost any growing problem ends up deeper than it first appears – but you have to be cognizant of this going in.

Unstructured Data > Structured

- Insights and data are valuable to businesses, but only when you have access to a source that the general market doesn't. The harder a valuable piece of data is to grab, the more attractive it is
- Many valuable pieces of data sit within unstructured text-based sources. It's notoriously tedious and difficult to extract insights from them
 - Ex: Public filings, public records PDF, transcriptions

Full-stack retrieval goes like this:

1. You have a raw corpus of documents (Held in the cloud)
2. You split them into semantically meaningful chunks (With LangChain or other text splitters)
3. You convert them into some vector representation for easy comparison and searching (Using OpenAI's embeddings)

4. You store those vectors (using [Pinecone](#) or [Weaviate](#) or [Chroma](#))
5. You retrieve certain documents based on the task at hand (Metal?)

I'm unsure how much of that stack a [Pinecone.io](#) is going to want to take vs a company like Metal – which is a current YC.

Contextualizing

We can't build unique models, but we can change the data through embeddings and update them affordably. Initial OpenAI embeddings, and cosine ranking is subpar after the initial wow factor. So to improve on models in private data, we need fine-tuning models with domain, incorporate keyword 'wut' search, and have multiple ranking methods.

Problem

- Semantic search gets you 90% of the way there for easy questions & answers, but only 30-40% for hard Q&A
- The hard part is understanding *which documents* are relevant to the query you give to the LLM

Why this is interesting to me

- I see two routes document retrieval could go
 - **Route #1 (Horizontal Retrieval):** One general engine is really good at document retrieval across industries and domains (Law, Medical, Real Estate, etc.). It has a reasoning engine that tells it where to look
 - **Route #2 (Verticalized Retrieval):** Specialized retrieval engines are needed who are experts at traversing law documents which are different than medical, real estate, etc.
- I'm unsure which way it will go! I'm currently leaning towards #2

The winner of this space will go full-stack and take over more document management / retrieval workflows

Vector Embeddings - learned matrix transformations that translate a dimensional space to another one while trying to go through a big information loss

Most places don't bother to define embeddings in general, or instead they describe the properties of the embeddings they want to use. Some want compression, some want cosine similarity.

At the end of the day, any medium that comes into these codes has to be converted to numerical vector. These conversions might be image converters, nlp text converters, audio converters etc. Not only do embeddings allow us to analyze and process vast amounts of data, but it also has the added benefit of being language-agnostic. Embeddings are modular independent and anything that's an input can be embedded.

The ability to vector match let's you have outcomes like "pink spiky fruit" mapping to dragonfruit instead of exact word matching that might lead to spiky fruit, or pink fruit etc... Put

easily, vector matching adds context. Basically even if you have things that don't have traditionally the same meaning, this will reduce it to points where the little amounts of nuance matter and we can match it to a specific part — meaning we can vector embed and capture meaning.

Given these vectors are essential, there is a need of databases for vectors that allow for storage, indexing, and servicing.

Vector search libraries help developers search through large collections of vectors for clusters or nearest neighbors. Popular ones include Google's [ScaNN](#) or Facebook's [Faiss](#). Vector search libraries are great for vector search, but they're not databases and have trouble at large scale.

Con

There are currently a thousand "load embedding vectors into a vector database and selectively load results into the context window" startups right now its crazy

Gap in Market

Pros

There's an issue with having an open source alternative that doesn't let users log in with GitHub and spin up an index and upload their vectors.

Features and Integration

One pain point that we noticed with a lot of existing vector stores is they often **involved connecting to an external server** that stored the embeddings. While that is fine for putting applications into production, it does make it a bit tricky to easily prototype applications locally.

They found that these were mostly geared to other use-cases and access patterns, like large-scale semantic search. Additionally, they were often a hassle to set up and run, especially in a development environment.

Since Chroma is deployed locally it will have lower latency than a managed cloud service due to network latency.

Cons

The issue with unmanaged — self hosted — vector databases:

- Self-hosted vector databases are a big step up from vector search libraries, but they still require significant configuration from engineering teams to scale without affecting latency or availability. They don't come with any **security guarantees** (i.e. GDPR or SOC 2 Type 2) and leave you with the **operational overhead** of maintaining additional infrastructure, monitoring additional services, and troubleshooting when things break. Solving these problems is where managed vector databases come into play.

Open Sourcing

Pro

- Can be the ability to move at the pace of AI. We don't know what it's bringing and the dimensional shifts it's going to take. So to keep up with the directional momentum of the ground moving underneath, letting users determine and help us evolve the databases might be a better way.
- AI is a landscape of shifting sands. What developers want today is not what they'll want in six months, and what they need to build demos is not what they need in production, is not what they'll need for integration into existing products. But demos *could* be the path to distribution.

Con

It's like anything else, **the risk adjusted returns are great enough to justify the most probable outcome..** at least in someone's book. Not all of these companies are being built to generate cashflows, at least a few are grinding until they can be acquired by someone who has a vision for how to extract value. Big fan of Langchain, fwiw.

The projects are highly technical so if a layperson wants to use it they pay \$\$\$ for a layperson dive into it

Another thing I suspect is if you get major corps to use your tech then their lives depend on your team so they'll "donate". This is actually tax deductible for them

Product Progression

Vector databases naturally sit at a critical point in the machine learning toolchain; any company with a lot of customers there would be well positioned to expand along that toolchain with new products. In particular, we can easily imagine a future where Pinecone begins offering a model hosting service, allowing them to manage the **entire vector data pipeline**.

Eventually, to win, the can become a truly seamless database for storing, indexing, and serving unstructured data. Bring your data:

- Vectorize
- Index
- Partition
- Store
- Query

Eventually, becoming an OLAP (online analytical processing) for unstructured data.

Every team wants to know the best way to leverage retrieval, how to chunk and embed their documents, which model they should use, how to ensure the retrieved data is relevant to the query — chroma will answer these questions

Bigger Fitting:

modular and flexible framework for developing AI-native applications.

"The real power comes when you are able to combine [LLMs] with other things."

LangChain aims to help with that by creating... a comprehensive collection of pieces you would ever want to combine... a flexible interface for combining pieces into a single comprehensive 'chain'

Edge over others;

Pessimism

Qdrant, Weaviate — clearly didn't market as well as Pinecone. James Briggs did an amazing job and he should be hottest DevRel in the space right now! Their blogs have high recall and the 'learn' series is often recommended

Hype Cycle and Market

The billion dollar question is whether all this interest leads to any durable market.

The cycle to date has been something like this:

- Models at scale have emergent behaviors that are magical and shocking.
- Consumers experienced DALLE2, ChatGPT, and a small number of LLM products gained real traction rapidly (Copilot, Jasper, Midjourney, Character).
- Startups have flocked to leverage these capabilities, VCs are funding them like it's 2021
- Many incumbent technology leadership teams are excitedly, anxiously resourcing AI projects.

Great companies can emerge from the morass of spaces like "LLMops." But those that do will be teams that see the wedge for what it is, rather than misreading immediate momentum and interest for durable value. The distance between Github stars and Twitter likes and at-scale deployments and six figure enterprise contracts is very far.

The question is not, "Do developers want LLMops?" but instead, "Which segment of those users do I focus on? What do they really need, and in what order? What will make the product easy to adopt, and what objections will I face? What architecture will support those users, and what compounding advantages can I build?" ((**** good hebbia starter email)

Current Worry Among Every Thinking Person

Too many people are hunting for a neat strategic narrative of "which layer of the stack endures," telling some clean story about "data moats," or wringing their hands that large labs or incumbents are going to win the core modalities (text, code, image etc.) — this kind of hand wringing is folly. the history of software markets is nondeterministic.

I believe, thee huge amount of value creation / capture out of the box for creative product folks is incredibly promising for startups. time and effort is better spent understanding customer problems deeply, and understanding the state of the art, and leveraging the latter for the former. who wins is based part on market structure, but also partly on who the players are, their execution, and how they redraw the software category lines

Intellectual Honesty Required: "thin shims on foundation model APIs" have fallen prey to technical arrogance. Copilot became quickly essential because co figured out how to fit

"passive" prediction into coding workflows in a way that made sense to developers. People building from the models/tools up (VS from the customer back) are often unwilling to focus enough to do that last mile to make a product useful for customers. Extreme amounts of CUSTOMER FUCKING CENTRICITY and building backwards.

- Think about whether it will matter for the use case once models improve
- Incorporate private data/customer data in the model context to improve outputs
- Assume that incumbents in your space will at least adopt surface-level generative AI features and think about how you can go beyond those.
 - Advantages of the incumbent:
 - Distribution
 - Prop. Data
 - Capital
 - Talent
 - Advantages of the startup
 - Speed
 - Focus
 - Centralization of data
 - Less repetitional risk
- Think about the right insertion point for your product and try to go deep into workflows while minimizing disruptions but bringing out the full value of AI.

Props to them

Developer Marketing: Clever move by chroma. Marketing is the key vector for DB companies' success, particularly for Vector DBs as we are still in the hacker/experimental phase.

Developers value familiarity and ease of use over technical features

I don't think the billions of LLM developers need to worry about scale. Chroma is moving vector infra out of data centers to the edge and your file system in an AI-first ecosystem. The reason why langchain hackers preferred chroma was that it was easy to use locally. Once you need to connect to an endpoint for scale, the complexity comes. There might be inability to scale...

Serving a customer's needs well – in this case usually developers and larger companies wanting to integrate AI to their systems – is often more important (and harder) to think about than defensibility. In many cases defensibility emerges over time - particularly if you build out a proprietary data set or become an ingrained workflow – which Chroma is likely to follow, or create defensibility via sales or other moats.

The less building and expansion of the product you do after launch, the more vulnerable you will be to other startups or incumbents eventually coming after and commoditizing you. Pace of

execution and ongoing shipping post v1 matters a lot to building one forms of defensibility above.

Well is DATA the new moat — building for these proprietary data sets???

Other thing to think about while servicing smaller customers on their ML Dev journey is the graduation issue – will they be too big to want to host it themselves, and can we scale alongside them (i.e the stripe phenomenon)

Enterprise document management

- By default, internal company documents (slides, docs, emails, messages, APIs) are not optimized for LLMs
- Big companies will need custom solutions to organize all of their internal documents for LLMs to parse and retrieve
- A company will emerge as "the first place for your LLMs to ingest your documents"
- [Unstructured](<https://www.unstructured.io/>) might be the front runner

▼ Early AI Meditations 3

Tweets on my mind

<https://twitter.com/dharmesh/status/1646584583637303296>

<https://twitter.com/sjwhitmore/status/1645811222661718021>

AI Trends I'm Interested In

- **Evolving Agent Roles - Specialized Agents**
 - **Notes**
 - Agents aren't as reliable as you'd like. They're better when they focus.
 - V2 Agent Frameworks and roles are starting to emerge. ReAct has been around but "stacking" together planner agents, executor agents, and overseeing agents are proving useful
 - **Hypothesis**
 - AutoGPT & Baby AGI are awesome! It's an early glimpse at the frameworks to come. However I'm still looking for the use case that demo's well (they all do) & provides value (haven't seen this yet)
 - **Questions on my mind**
 - What's the optimal role configuration and number of agents to complete a task? Will it be modeled after human best practices, or new ones? General? Verticalized?

- AI simulation w/ the 25 agents was super interesting. I wonder what sort of hive mind theory you get with a bunch of them interacting with each other.
- How long will it be till best practices evolve to the point where we trust production grade agents money & APIs?
- **LLM per person per task (Tone Personalization)**
 - **Notes**
 - Izzy Miller showed how he trained an LLM on 500K messages group messages between 5 people. He mentioned that he did not train 5 LLMs, but rather one. The result was a bot that could respond back in 'group' tone.
 - **Hypothesis**
 - It's a no brainer that each person will have an LLM trained on their style, tone, and preferences. However you wear different hats in life. Work. Play. Social. You'll need a system that understands each.
 - **Open Questions**
 - Will an individual want LLMs to impersonate them in the long term? Or will I point people towards "Greg's AI" instead of "Greg"?
 - Will you own your LLM? Or will each application will have an LLM that knows you? I'm leaning towards a mix of the two
- **AI Fault Tolerance**
 - **Notes**
 - Agents aren't perfect: Sam Whitmore (top LangChain contributor) estimates them to be successful 85% of the time.
 - As more agents handle decision making, frameworks and tech to support these non-deterministic outputs will be critical
 - **Hypothesis**
 - This will likely happen on the application framework level (aka LangChain)
 - AI Eval will be extremely important - companies will be built here. Example
- **Idea generation via AI**
 - **Notes**
 - It's an aspirational goal to have AI create new knowledge for us. Specially in business & scientific research.
 - I'm a fan of David Deutsch's explanation and criticism.
 - **Hypothesis**

- There will be an awesome framework coming out shortly that specialized in idea generation through generating hypotheses and then self-evaluating them or building out proof like logic. Similar to BabyAGI and AutoGPT
- **On my to do list**
 - Editing BabyAGI to focus on idea generation and criticism. Instead of "make me a marketing plan" I'd love to see an optimized BabyAGI tackle, "here is this unsolved problem in math, iterate on it until you have an answer. Grade yourself"
- **Enterprise document management**
 - **Notes**
 - By default, internal company documents (slides, docs, emails, messages, APIs) are not optimized for LLMs
 - Big companies will need custom solutions to organize all of their internal documents for LLMs to parse and retrieve
 - A company will emerge as "the first place for your LLMs to ingest your documents"
 - **Hypothesis**
 - Unstructured is the front runner

▼ Early AI Meditations 2

Tweet on my mind

<https://twitter.com/karpathy/status/1642607620673634304>

AI Trends I'm Interested In

- **Managed Retrieval Engines**
 - **Problem**
 - Semantic search gets you 90% of the way there for easy questions & answers, but only 30-40% for hard Q&A
 - The hard part is understanding *which documents* are relevant to the query you give to the LLM
 - **Why this is interesting to me**
 - I see two routes document retrieval could go
 - **Route #1 (Horizontal Retrieval):** One general engine is really good at document retrieval across industries and domains (Law, Medical, Real Estate, etc.). It has a reasoning engine that tells it where to look

- **Route #2 (Verticalized Retrieval):** Specialized retrieval engines are needed who are experts at traversing law documents which are different than medical, real estate, etc.
- I'm unsure which way it will go! I'm currently leaning towards #2

- **Notes**

- Metal (Managed Retrieval) just announced an [integration with LangChain](#)
- This topic likely deserves its own essay in the future. Here's the TLDR of that essay already:
 - Full-stack retrieval goes like this:
 1. You have a raw corpus of documents (Held in the cloud)
 2. You split them into semantically meaningful chunks (With LangChain or other text splitters)
 3. You convert them into some vector representation for easy comparison and searching (Using [OpenAI's embeddings](#))
 4. You store those vectors (using [Pinecone](#) or [Weaviate](#))
 5. You retrieve certain documents based on the task at hand (Metal?)
 - I'm unsure how much of that stack a [Pinecone.io](#) is going to want to take vs a company like Metal.

- **Hypothesis**

- The winner of this space will go full-stack and take over more document management / retrieval workflows

- **Developer Monetization with OpenAI Plugins**

- **Problem**

- In a market place you need adoption incentives on both sides to drive overall health. Without monetization for the plugin supply side (developers) it's hard to get the demand
- OpenAI has [100M+ users](#), now they need to incentivize developers to build & maintain plugins

- **Hypothesis**

- There *might* be a use case for micro-transactions (very hesitant to use that word) for plugin use that happens through OpenAI
- LLM Plugin access will become a standard feature line on pricing tiers for virtually every company. Starting at the top (enterprise/mid-market) and working its way down to SMBs as more SMB-friendly tools get built

- **Notes:**
 - I shutter at the words 'micro-transactions' because with all the talk over the past few years we have yet to see them happen in a material way
 - Plugin user level auth will make this seamless
- **Plugin Translators Dev Shops**
 - **Problem**
 - Businesses will want their services to be accessible to LLMs, but they won't all have the skills required to create, maintain, and develop plugins
 - **Hypothesis**
 - There will be shops that specialize in creating and maintaining plugins for companies. A small dev shop could likely 'translate' thousands of APIs at a time
 - Monetization incentives (above) will drive this
 - PSO (Plugin Store Optimization) will evolve out of too much supply
 - **Notes**
 - An early look at what this world will look like:
 - <https://twitter.com/matchaman11/status/1641502642219388928>
- **Unstructured Data > Structured**
 - **Problem**
 - Insights and data are valuable to businesses, but only when you have access to a source that the general market doesn't. The harder a valuable piece of data is to grab, the more attractive it is
 - Many valuable pieces of data sit within unstructured text-based sources. It's notoriously tedious and difficult to extract insights from them
 - Ex: Public filings, public records PDF, transcriptions
 - **Hypothesis**
 - There will be an addition to the data-service industry (like CBInsights) enabled by LLMs. BUT you won't hear about it because suppliers know that their data's value is derived from its scarcity. It's not in their best interests to tell you how it's gathered
 - **Examples**
 - **Tech Extraction from Job Descriptions**

- Sales teams often use job descriptions to parse which technologies or tools a company is using. This is usually a manual process. It's like BuiltWith but for job descriptions. With LLMs it's easy to parse JDs to get this information.
- (Disclosure: I built this)

<https://twitter.com/GregKamradt/status/1643027796850253824>

- **Community Moderation & Analytics (Discord/Slack/Support)**

- **Analytics:** You have better ways to classify and report on conversations & requests in your community. Businesses would 100% pay for this if you give recommendations on how to increase health.
- **Moderation:** Users post questions to the wrong channel. It would be nice to clean those up by going through them, classifying them, and moving them. Or stopping users from posting them all together

- **Reflection**

- **Problem**

- LLMs are good, but not always on their first draft of a response

- **Solution**

- It's super easy to ask them, "are you sure?" and get a better answer back. It's been statistically proven to increase quality of answers over a varying level of benchmarks

- **Notes**

- Unfortunately reflection increases costs and latency since you're making another API call. This isn't a problem for all use cases, but users can be time-insensitive.

- **Resource: Great Video on the topic**

<https://www.youtube.com/watch?v=5SgJKZLBrmg>

- **Drag & Drop LLM/Chain Builders**

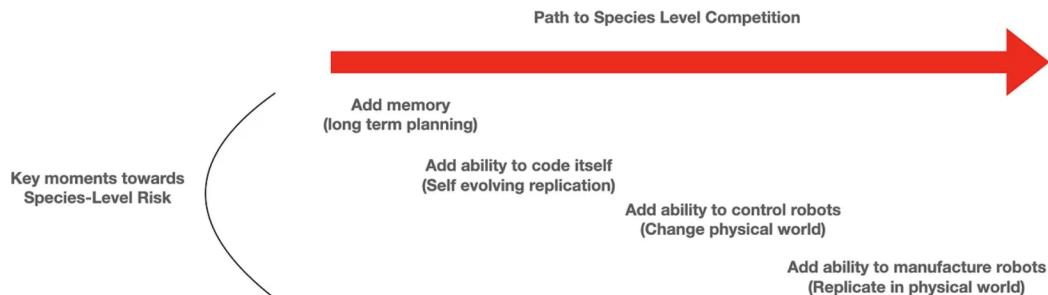
- **Problem**

- Not everyone is technical. Even if you are, it's sometimes easier to drag and drop rectangles on a screen than write code

- **Opportunity**

- Create no code tools that string together LLM calls

- Basically no-code [LangChain](#)
- Notes
 - Examples: [LangFlow](#)
- Hypothesis
 - I don't think this will be as big as it may seem. Simple no-code use cases are easier (aka Zapier), but going deep requires technical ability (aka Bubble/webflow)
 - Opinion: It's interesting eye-candy but I wouldn't recommend investing here
- Elad Gil: Species Level Take Over - [Link](#)
 - Notes
 - This is just thought-candy, but I thought Elad had an interesting framework to think about different tiers
 - "For AI to move from **merely another technology risk** (in the long line of tech risks we have survived and benefited from on net) to a potentially **existential species-level risk** (all humans can die from this), up to two technological breakthroughs need to happen (and (2) below - robotics, may be sufficient):"
 1. The AI needs to start coding itself and evolve: tool → digital life transition.
 2. Robotics need to advance: Digital→real world of atoms transition.
 - Species Level Competition



▼ Neural Networks from Scratch

Nowadays, creating an AI algorithm is incredibly easy. Using a few libraries will allow any model to be created with only a few lines of code.

But I realized that this made me incredibly disconnected with the underlying principles of AI. The math: the calculus, linear algebra, and probability that is the backbone of neural networks. So to get back in touch with the fundamentals, I built a vanilla neural network using nothing but python. No libraries, just math.

ANN from scratch no libraries.py

```
# import libraries
import csv
import math
import random

# Import dataset
with open("/Users/alexander/Anaconda Projects/ANN from scratch - no libraries.csv") as csvfile:
    dataset = list(csv.reader(csvfile))

# Delete label row
del dataset[0]
# Drop unnecessary columns
for row in range(len(dataset)):
    del dataset[row][0:3]
#Shuffle the dataset
random.shuffle(dataset)

#one hot encoding
def one_hot_encode (dataset, column):
    rows = len(dataset)
    things = []
    for row in range(rows):
        things.append(dataset[row][column])
    items = []
    items.append(dataset[0][column])
    for row in range(rows):
        thing = dataset[row][column]
        if thing not in items:
            items.append(dataset[row][column])
    for item in items:
        for row in range(rows):
            if things[row] == item:
                dataset[row].insert(column, 1)
            else:
                dataset[row].insert(column, 0)
    for row in range(len(dataset)):
        del dataset[row][column+len(items)]
    return dataset
```

```

country_dataset_ohe = one_hot_encode(dataset = dataset, column = 1)
dataset = one_hot_encode(dataset = country_dataset_ohe, column = 4)

# Making sure all vlaues are doubles
for row in range(len(dataset)):
    dataset[row] = list(map(float, dataset[row]))


# Split dataset into test and train
num_rows = len(dataset)
split = math.floor(num_rows*0.7)
dataset_train = dataset[:split][ :]
dataset_test = dataset[split:][ :]

# Split dataset into X and Y values
num_col = len(dataset[0])
spliting = num_col - 1
dataset_X_train = []
dataset_X_test = []
dataset_Y_train = []
dataset_Y_test = []
for row in range(len(dataset_train)):
    dataset_X_train.append(dataset_train[row][:spliting])
    dataset_Y_train.append(dataset_train[row][spliting:])
for row in range(len(dataset_test)):
    dataset_X_test.append(dataset_test[row][:spliting])
    dataset_Y_test.append(dataset_test[row][spliting:])


# Feature Scaling
def feature_scale_set (dataset):
    mean = []
    for col in range(len(dataset[0])):
        summation = 0
        for row in range(len(dataset)):
            summation = summation + dataset[row][col]
        tempmean = summation / len(dataset)
        mean.append(tempmean)
    maxmin = []
    for col in range(len(dataset[0])):
        largest = dataset[0][col]
        smallest = dataset[0][col]

```

```

        for row in range(len(dataset)):
            if dataset[row][col] > largest:
                largest = dataset[row][col]
            if dataset[row][col] < smallest:
                smallest = dataset[row][col]
        tempmaxmin = largest-smallest
        maxmin.append(tempmaxmin)
    return mean, maxmin

def feature_scale (dataset, mean, maxmin):
    num_col = len(dataset[0])
    num_row = len(dataset)
    for col in range(num_col):
        for row in range(num_row):
            dataset[row][col] = (dataset[row][col] - mean[col])/maxmin[col]
    return dataset

mean, maxmin = feature_scale_set(dataset_X_train)
dataset_X_train = feature_scale(dataset_X_train, mean, maxmin)
dataset_X_test = feature_scale(dataset_X_test, mean, maxmin)

# Make each example a column instead of row
def transpose (dataset):
    newdataset = []
    numrows = len(dataset)
    numcols = len(dataset[0])
    for col in range(numcols):
        temprow = []
        for row in range(numrows):
            temprow.append(dataset[row][col])
        newdataset.append(temprow)
    return newdataset

X_train = transpose(dataset_X_train)
X_test = transpose(dataset_X_test)
Y_train = transpose(dataset_Y_train)
Y_test = transpose(dataset_Y_test)
'''

X_train = dataset_X_train
X_test = dataset_X_test
Y_train = dataset_Y_train
Y_test = dataset_Y_test

```

```

'''



# The dimensions of each layer
layer_dim = [ len(X_train), 6, 6, 1]

#Initializing parameters
def init_parameters(dim):
    parameters = {}
    num_layers = len(dim)
    for layer in range(1, num_layers):
        matrix = []
        bmatrix = []
        for row in range(dim[layer]):
            temprow = []
            for col in range(dim[layer-1]):
                temprow.append(random.random()*0.1)
            matrix.append(temprow)
            tempzero = [0.0]
            bmatrix.append(tempzero)
        parameters['W' + str(layer)] = matrix
        parameters['b' + str(layer)] = bmatrix
    return parameters

# Calculate the dot product
def dot(matrix1, matrix2):
    matrix = []
    col2 = len(matrix2[0])
    row1 = len(matrix1)
    dims = len(matrix2)
    if dims == len(matrix1[0]):
        for row in range(row1):
            temprow = []
            for col in range(col2):
                value = 0
                for dim in range(dims):
                    value = value + (matrix1[row][dim] * matrix2[dim][col])
                temprow.append(value)
            matrix.append(temprow)
        return matrix
    else:
        print("The column of first matrix doesn't match with row of second")

#add the bias matrix element-wise

```

```

def addb(matrix, b):
    newmatrix = []
    numrows = len(b)
    numcols = len(matrix[0])
    if numrows == len(b):
        for row in range(numrows):
            temprow = []
            for col in range(numcols):
                temprow.append(matrix[row][col] + b[row][0])
            newmatrix.append(temprow)
        return newmatrix
    else:
        print("The dimensions don't match. The column of each column doesn't match the number of rows in matrix")
        return None

### this is to do sigmoid activation
def sigmoid(matrix):
    newmatrix = []
    numrows = len(matrix)
    numcols = len(matrix[0])
    for row in range(numrows):
        temprow = []
        for col in range(numcols):
            value = matrix[row][col] * -1
            value2 = math.exp(value)
            value3 = 1 + value2
            value4 = 1/value3
            temprow.append(value4)
        newmatrix.append(temprow)
    return newmatrix

# Multiply element-wise
def multiply(matrix, num):
    newmatrix = []
    numrows = len(matrix)
    numcols = len(matrix[0])
    for row in range(numrows):
        temprow = []
        for col in range(numcols):
            temprow.append(matrix[row][col] * num)
        newmatrix.append(temprow)
    return newmatrix

# Calculate exponents for each element

```

```

def exp (matrix):
    newmatrix = []
    rows = len(matrix)
    cols = len(matrix[0])
    for row in range(rows):
        temprow=[]
        for col in range(cols):
            temprow.append(math.exp(matrix[row][col]))
        newmatrix.append(temprow)
    return newmatrix

# Forward propagation up until before output layer
def forward(X, parameters, dim):
    cache = {}
    cache['A0'] = X
    for layer in range(1, len(dim)-1):
        Zpre = dot(parameters['W' + str(layer)], cache['A' + str(layer-1)])
        Z = addb(Zpre, parameters['b' + str(layer)])
        cache['Z' + str(layer)] = Z
        A = []
        for row in range(len(Z)):
            temprow = []
            for col in range(len(Z[0])):
                if Z[row][col] < 0:
                    temprow.append(0)
                else:
                    temprow.append(exp(Z[row][col]))
            A.append(temprow)
        cache['A' + str(layer)] = A
    return cache

#Now do the output layer
def output(cache, parameters, dim):
    layer = len(dim) -1
    Zpre = dot(parameters['W' + str(layer)], cache['A' + str(layer-1)])
    Z = addb(Zpre, parameters['b' + str(layer)])
    cache['Z' + str(layer)] = Z
    A = sigmoid(Z)
    cache['A' + str(layer)] = A
    return cache

```

```

# Cost function
def cost_function(Y, cache, dim):
    output = cache['A' + str(len(dim)-1)]
    m = len(Y)
    if m == len(output):
        loss = 0
        for example in range(m):
            loss = loss + ((Y[0][example] * math.log(output[0][example])))
        cost = (loss/m) * - 1
        return cost
    else:
        print("The number of examples for the ouput and the answer doesn't
# Gradient for output
def get_gradients(Y, parameters, cache, dim):
    gradients = {}
    # Find final layer dZ
    temprow = []
    dzfinal = []
    for example in range(len(Y[0])):
        temprow.append(cache['A' + str(len(dim) - 1)][0][example] - Y[0][example])
    dzfinal.append(temprow)
    gradients['dZ' + str(len(dim) - 1)] = dzfinal
    # Find final layer dw
    gradients['dw' + str(len(dim) - 1)] = multiply(dot(gradients['dZ'+ str(len(dim) - 1)], parameters['W' + str(len(dim) - 1)]))

    # Find final layer dB
    db = []
    dbpre= []
    bsum = 0
    for example in range(len(Y[0])):
        bsum = bsum + gradients['dZ' + str(len(dim) - 1)][0][example]
    dbpre.append(bsum/len(Y[0]))
    db.append(dbpre)
    gradients['db' + str(len(dim) - 1)] = db
    # For all subsequent layers
    for layer in range(len(dim)-2,0,-1):
        dZpre = dot(transpose(parameters['W' + str(layer+1)]), gradients['dZ' + str(len(dim) - 1)])
        dZ = []
        for row in range(len(dZpre)):
            temprow = []
            for col in range(len(dZpre[0])):
                value = cache['Z' + str(layer)][row][col]
                temprow.append(value * gradients['dZ' + str(len(dim) - 1)][row][col])
            dZ.append(temprow)
        gradients['dZ' + str(layer)] = dZ

```

```

        if value < 0:
            temprow.append(0)
        else:
            temprow.append(value)
        dZ.append(temprow)
    gradients['dZ' + str(layer)] = dZ

    gradients['dW' + str(layer)] = multiply(dot(gradients['dZ'+ str(la
        db = []
        for row in range(len(gradients['dZ' + str(layer)])):
            rowsum = 0
            bsum = []
            for col in range(len(gradients['dZ' + str(layer)][0])):
                rowsum = rowsum + gradients['dZ' + str(layer)][row][col]
            bsum.append(rowsum/len(Y[0]))
            db.append(bsum)
        gradients['db' + str(layer)] = db

    return gradients

# Element wise subtraction
def subtract(matrix1, matrix2):
    newmatrix= []
    numrows = len(matrix1)
    numcols = len(matrix1[0])
    if numrows == len(matrix2) and numcols == len(matrix2[0]):
        for row in range(numrows):
            temprow = []
            for col in range(numcols):
                value = matrix1[row][col] - matrix2[row][col]
                temprow.append(value)
            newmatrix.append(temprow)
        return newmatrix
    else:
        "The dimensions of your matrices do not match"

# Gradient descent
def gradient_descent(parameters, gradients, dim):
    for layer in range(1, len(dim)):
        parameters['W' + str(layer)] = subtract(parameters['W' + str(layer)], gradients['dW' + str(layer)])
        parameters['b' + str(layer)] = subtract(parameters['b' + str(layer)], gradients['db' + str(layer)])
    return parameters

```

```

parameters = init_parameters(layer_dim)
for x in range(50):
    print('Gradient descent step number: ' + str(x))
    cache = forward(X_train, parameters, layer_dim)
    cache = output(cache, parameters, layer_dim)
    cost = cost_function(Y_train, cache, layer_dim)
    print(cost)
    gradients = get_gradients(Y_train, parameters, cache, layer_dim)
    parameters = gradient_descent(parameters, gradients, layer_dim)

answer = Y_train[0]
total = len(answer)
guess = []
preguess = cache['A' + str(len(layer_dim)-1)][0]
numcols = len(cache['A' + str(len(layer_dim)-1)][0])
for col in range(numcols):
    if pregues [col] < 0.5:
        guess.append(0)
    else:
        guess.append(1)
right = 0
wrong = 0
for example in range(total):
    if answer[example] == guess[example]:
        right = right + 1
    else:
        wrong = wrong + 1
print()
print("You got " + str( 100 * (right/total)) + "% right on the training set")
print("Your final error is " + str( cost))
print()

def result(X_test, Y_test, parameters, layer_dim):
    cachetest = forward(X_test, parameters, layer_dim)
    cachetest = output(cachetest, parameters, layer_dim)
    cost = cost_function(Y_test, cachetest, layer_dim)
    total = len(Y_test[0])
    answer = Y_test[0]
    pregues = cachetest['A' + str(len(layer_dim)-1)][0]

    guess = []
    numcols = len(cachetest['A' + str(len(layer_dim)-1)][0])

```

```

for col in range(numcols):
    if preguess [col] < 0.5:
        guess.append(0)
    else:
        guess.append(1)
right = 0
wrong = 0
for example in range(total):
    if answer[example] == guess[example]:
        right = right + 1
    else:
        wrong = wrong + 1
print("You got " + str( 100 * (right/total)) + "% right on the test set")
print("Your test error is " + str( cost))

result(X_test, Y_test, parameters, layer_dim)

```

▼ Early AI Meditations 1

Tweet on my mind

<https://twitter.com/blader/status/1640387925912477698>

AI Trends I'm Interested In

-  **AI Memory** - LLMs are great reasoning engines, not great at memory. Major opportunity for players to provide infra to help with this. Likely will be verticalized
 - **Problem**
 - In-context learning works, however you need to elegantly select the right context you'd like your model to have.
 - Similarity search only goes so far. Most solutions only do top-N results. Lack of connecting ideas.
 - **Solutions today**
 - Similarity search via Pinecone, Weaviate, etc.
 - **Hypothesis**
 - Different verticals will need different knowledge graph expertise. Law vs medical vs sales vs product vs user research. Verticalized players will likely emerge
 - **Notes:**

- **OpenAI mentions better memory on their plugin's next steps** - "Integrating more optional services, such as summarizing documents or pre-processing documents before embedding them, could enhance the plugin's functionality and quality of retrieved results. These services could be implemented using language models and integrated directly into the plugin, rather than just being available in the scripts."
-  **LLM Coordinators (ex: LangChain)** - Organizing, customizing and providing modularity to LLM applications
 - **Problem**
 - Developers need ways to customize how their product consumes and instructs language models.
 - All developers run through the same friction when building apps. Prompt templating, retries, parsing output.
 - **Solution today**
 - Build your own ergonomics
 - Use a library ([LangChain](#), [LLama Index](#))
 - **Notes**
 - Libraries like LangChain make it easier to work with LMMs. It's unclear how much OpenAI and other companies will strategically build products into the space. Ex: LangChain and LLama index are great at document loading. Developers now need to choose if they load docs through them or use an OpenAI Plugin.
 - Model swapping, finer tuned control over agents, definitely needed.
-  **Internal company APIs** - Proprietary Plugins for internal company use
 - **Notes**
 - Plugins could be a beautiful way for LLMs to chat with external facing apps. A cute and demo worth example of this is ChatGPT booking a [dinner reservation](#).
 - **Hypothesis**
 - My hypothesis is that companies will have an internal LLM that carries out instructions with internal facing apps and plugins.
 - While large enterprise might do this themselves to start, my hypothesis is that Mid market/SMB will outsource this to products that do it for them
 - **Example applications**
 - Some companies are so massive that it's difficult to know what is going on around the org. It would be great if there was an LLM that was watching a feed and only alerted me of what I needed
 - Trained specifically on a company's code base and could make recommendations

- Could train product marketing to better articulate how code works
 - Keep up technical documentation up to date
 - This will be similar
-  **No code ways to make your own apps** - Big opportunity to empower people to make their own apps powered by AI.
 - **Problem**
 - Non-technical people have great ideas, but can't build apps to execute them
 - **Hypothesis**
 - Low-code and no-code has already been around, but the barrier to entry is still too high. As english becomes a programming language more SMB owners will build apps that have a solid use case
 - Micro-SaaS acquisition could likely heat up here. If not to purchase a company, then for start up that can execute better to run with their idea.
-  **Offshoots of Plugin Store** - Apple AI App Store
 - **Notes**
 - OpenAI decided to use an open API specification format, where every API provider hosts a text file on their website saying how to use their API.
 - This means even this plugin ecosystem isn't a closed off that only a first mover controls
 - **Hypothesis**
 - Most of the infrastructure and support we see around the apple app store will likely follow the plugin store
-  **LLM Privacy** - The Signal of LLMs
 - **Notes**
 - The company to crack a private LLM (Ex: Get the reasoning power of an LLM but with complete privacy) will gain massive traction.
 - **Hypothesis**
 - This is a horizontal feature that would likely be extremely attractive to OpenAI and other providers
 - Can we reduce the security threats present in the way we treat llms

▼ Assessing the Influence of Neutral Grounds on Match Outcomes

Running head: ASSESSING THE INFLUENCE OF NEUTRAL GROUNDS

`Assessing the influence of neutral grounds on match outcomes`

Amenti Kenea

Executive Summary

Kneafsey and Müller's paper highlights the positive impact of neutral grounds on sports leagues by providing a balancing effect on home advantage (Kneafsey & Müller, 2018). This memo replicates these findings and extends them by using genetic matching to estimate the causal effect of neutral grounds on match outcomes and conducts a sensitivity analysis to check the robustness of its results. Since it finds that neutral grounds indeed provide a fairer playing field in terms of location-based advantages between teams, this memo recommends the increased use of neutral grounds by the Gaelic football and hurling associations to foster fairer competition while reducing monetary, psychological and physiological costs for teams and players alike. It also suggests further research should be carried out to extend the findings to other sporting events such as international leagues.

The advantage of playing at home is a well-known phenomenon in sports. It is a well-researched fact that teams playing at home have an advantage over the opposing team (Kneafsey & Müller, 2018; Carroll & Collins, 2012; Pollard, 2008; Pollard & Pollard, 2005; Pollard, Prieto, & Gómez, 2017). Many intuitive reasons have been put forward: teams feel more comfortable in their own turf since they practice regularly on it, players experience less performance anxiety due to the familiarity of the environment, the climate is not different from that in which they usually train, the home crowd of fans is more supportive, and so on. It is not difficult to see why the team playing at home would have a considerable advantage over the team that is visiting, in almost all team sports.

The use of neutral grounds, however, although intuitively a "middle ground" between the home and away fields, is not widely used. Many leagues and championships around the world use a home-away dichotomy, where teams play one match at home and one match away and the aggregate result of both games determines who advances to later rounds (for example, the UEFA Champions League [UEFA, 2019]). Neutral grounds are most often used for final games, where the outcome of the match determines the winner of the championship, but these occur very sparingly, and cannot be directly compared to other games since the teams do not have the possibility to advance any further in a final game -- the motivations are different. Thus, although neutral grounds seem like an intuitive option to provide a counterbalance for the home advantage without having to play two games in two different locations, it is not widely employed except for exceptional circumstances such as a final game.

In their paper, Liam Kneafsey and Stefan Müller (Kneafsey & Müller, 2018) examine the question of whether neutral grounds provide the desired counterbalancing effect by looking at two specific sports leagues, Gaelic football and Gaelic hurling, in Ireland. They find that neutral grounds indeed provide a balance, taking away from the home advantage and giving an extra edge to the away team. However, by using coarsened exact matching they are only able to use

a fraction of the original dataset, which means results could be improved upon by using more advanced matching methods. Additionally, they don't match on a large number of covariates -- thus, a sensitivity analysis could provide insight into how robust the results actually are. This paper sets out to replicate the results in Kneafsey and Müller (Kneafsey & Müller, 2018) and extend them by using genetic matching and then conducting a Rosenbaum sensitivity test on the results, along with a discussion highlighting the potential benefits of using neutral grounds

The homogeneity of Gaelic football and hurling in Ireland provide a controlled environment to estimate the causal effect of neutral grounds

The main reason for choosing Gaelic football and hurling as the sport to analyse in Kneafsey and Müller's paper is the homogeneity of the sport in the country and the volume of games played. Ireland is not very large, so there is little (if any) variation in natural factors like weather or temperature that could have an impact on different teams within Ireland. Furthermore, Gaelic football is an amateur league, meaning that the fields these teams play in are fairly similar to each other, since there are no large disparities between the wealth of different clubs. Gaelic football is also by far the best attended sport in Ireland, with hurling being the third behind association football (Kneafsey & Müller, 2018), meaning that there is a lot of data to work with. The dataset the paper uses is freely available online and the results can thus be replicated by anyone.

Although only focusing on Gaelic football and hurling makes the results of the paper less generalizable to other sports around the world, it allows for the natural control of many confounding variables that would otherwise be very hard to control, as mentioned in the previous section. The homogeneity of Ireland and the sports leagues themselves compared to international sports leagues provide a more accurate assessment of the effect of neutral grounds on match outcomes.

Kneafsey and Müller (Kneafsey & Müller, 2018), and this paper, analyze all 3,503 competitive senior men's inter-county hurling and Gaelic football matches between January 2009 and June 2018. To understand the impact of neutral grounds, the analysis focuses on the favoured team only (based on Elo[1] ratings). Furthermore, the analysis incorporates information about the relative team strength (using Elo ratings) and the importance of the match (Championship, latestage games, and others) to control for factors that could influence the final outcome of the game. The "treatment" of the dataset, therefore, is whether the game was played on a neutral ground or not, with the "outcome" being the win/loss of the team, and the "dependent" variables being the Elo rating and the importance of the match (along with, of course, which sport is being analyzed).

Our replication verifies the results obtained by Kneafsey and Müller and illustrates the balancing impact of neutral grounds on match outcomes for the home and away teams

We make use of the source code provided by Kneafsey and Müller to replicate their results. By carrying out the same preprocessing procedures, we obtain *Figure 1* below, which shows the bootstrapped mean results of the outcome of games played at home, away, and on neutral

grounds for the favorite team (dictated by Elo rating) by sport. A 1 indicates a win and 0 indicates a loss.

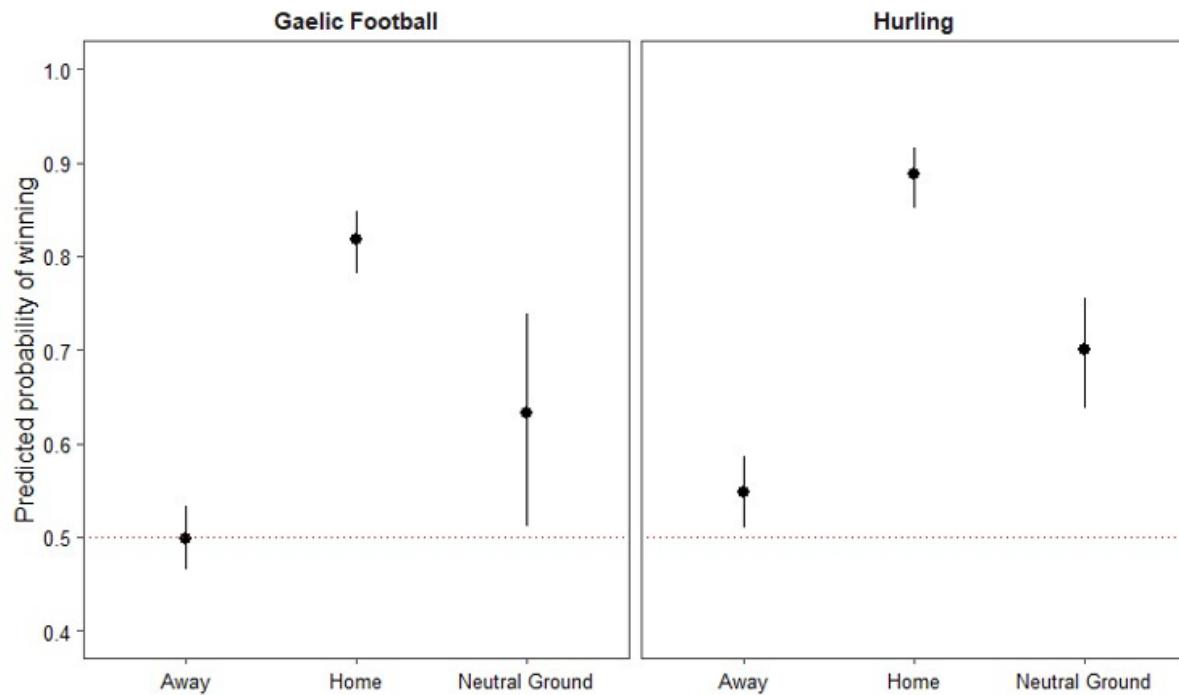


Figure 1. Bootstrapped means of the predicted probability of winning based on the type of ground being played at, by sport.

This is purely descriptive, however -- it does not tell us anything about the causal effect of playing on neutral grounds. For that, we carry out coarsened exact matching using the same procedures that Kneafsey and Müller employed, and obtain the results shown in *Figure 2*. The figure is the same that was obtained in the original paper.

The results obtained by Kneafsey and Müller indicate that neutral grounds do indeed have a balancing impact on the advantage of the home team, with the log-likelihood of the favorite team increasing by 0.798 when playing on a neutral ground instead of an away field and decreasing by -1.037 when playing on a neutral ground instead of a home field. However, Kneafsey and Müller are only able to match on 37% of the away data, and 54% of the home data, meaning that this effect represents the sample average treatment effect on the treated (SATT) for the data points that they were able to match on. This makes their results less generalizable, and potentially less robust, as opposed to using the entire dataset.

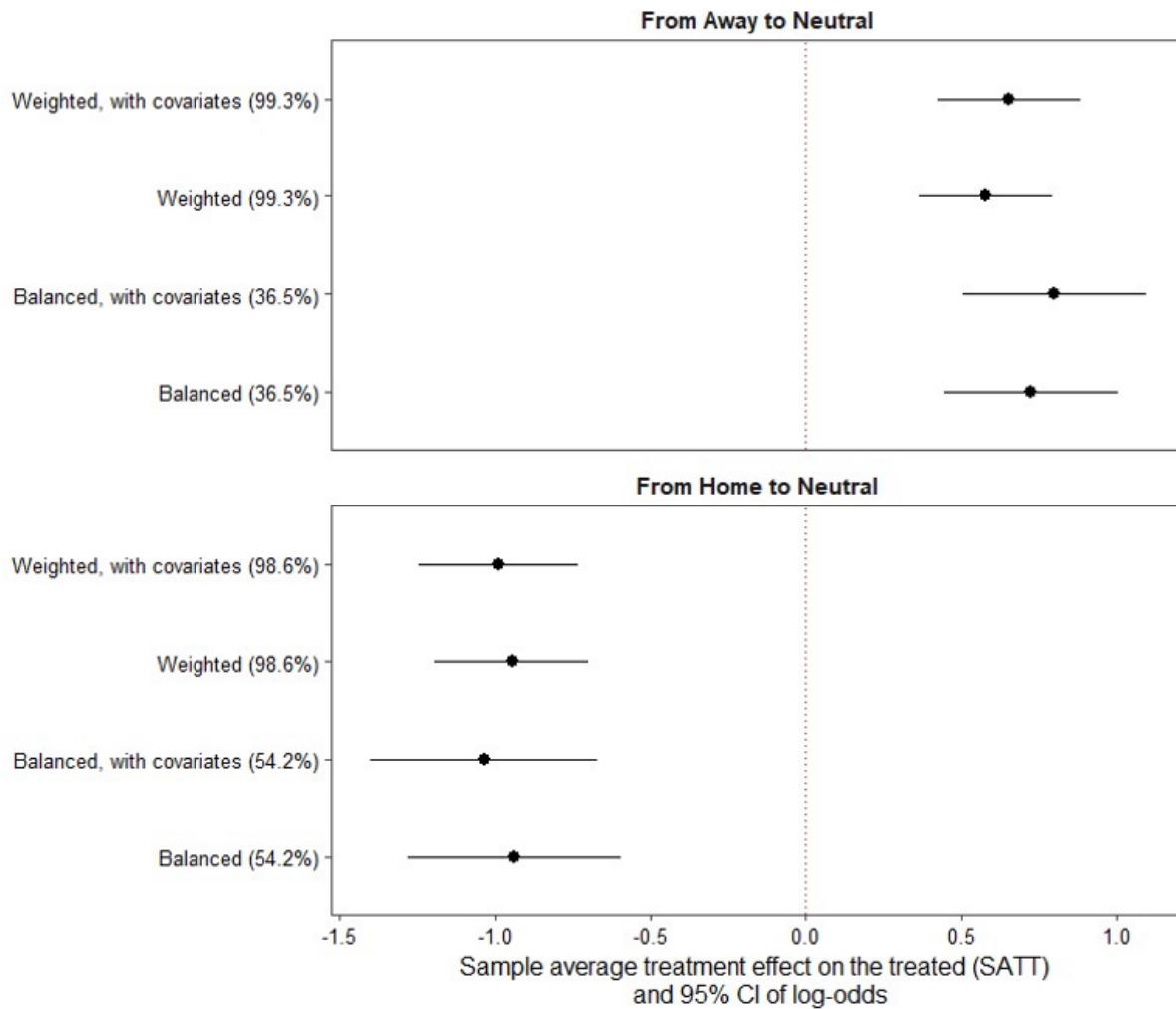


Figure 2. Causal effect of using neutral grounds on the log-odds of winning a match for the favoured team when the control is playing away and playing at home, respectively. Weighted results show the causal effect without coarsened exact matching (i.e. multivariate regression), while balanced results show the effect after conducting coarsened exact matching. With covariates refers to the inclusion of covariates in the final logistic regression model used to estimate effects. Percentages refer to the percentage of the dataset used. We see that the log-odds of winning decrease when going from home to neutral and increase when going from away to neutral.

Using genetic matching instead of coarsened exact matching adds robustness to the results obtained by increasing the sample size used

In this paper, we extend this analysis by carrying out genetic matching instead of coarsened exact matching. Through genetic matching, we are able to retain 97.8% of the away game data and 99.6% of the home game data, thus adding robustness to the results obtained by Kneafsey and Müller and also improving the generalizability of the results, as the sample examined is now the entire Gaelic football and hurling leagues in the dataset instead of only a fraction of it.

Nevertheless, the results we obtain using genetic matching for the causal effect of playing on neutral grounds on the probability of winning for the favored team are very similar to the results obtained by Kneafsey and Müller originally, with the log-likelihood of the favorite team increasing by 0.769 when playing on a neutral ground instead of an away field, and decreasing by -0.904 when playing on a neutral ground instead of a home field. They are displayed in *Figure 3*, alongside the results from the original paper.

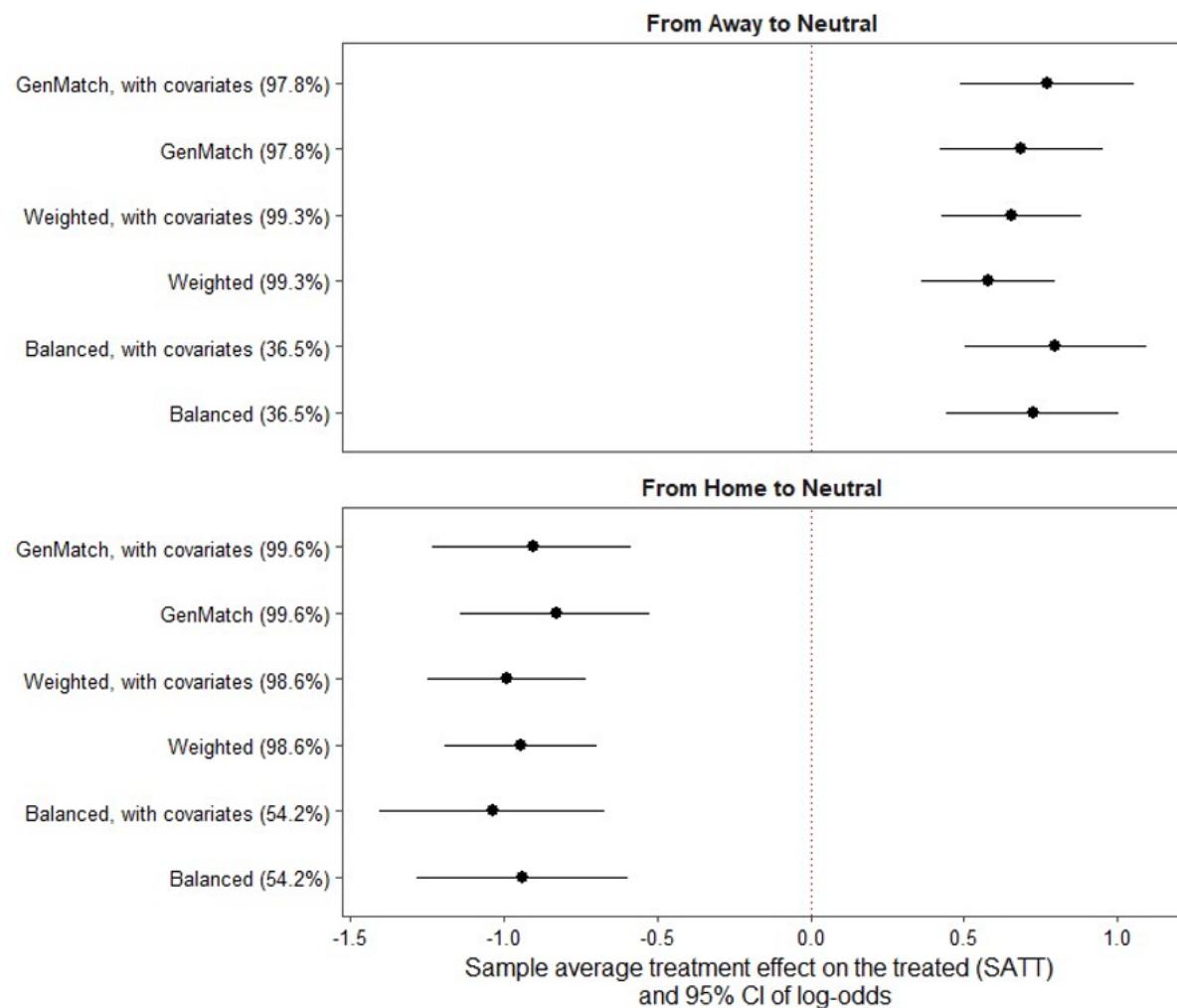


Figure 3. Causal effect of using neutral grounds on the log-odds of winning a match for the favoured team when the control is playing away and playing at home, respectively. This figure is an extension of *Figure 2*, where GenMatch indicates the results of using genetic matching both with and without including covariates in the final logistic regression model used to evaluate causal effects. We see the results are very similar to those obtained by Kneafsey and Müller.

Conducting a Rosenbaum sensitivity analysis demonstrates the results obtained are relatively sensitive to hidden bias

This paper also extends the results obtained by Kneafsey and Müller by carrying out a sensitivity analysis. Originally, Kneafsey and Müller control for only three covariates that they consider of importance: the relative team strength, measured by the Elo rating, the level of importance of the match, and naturally the type of sport. They justify that other covariates likely have a negligible impact due to the homogeneity of Ireland as a country and the Gaelic leagues as a whole. However, this paper conducts a Rosenbaum sensitivity analysis to investigate whether an unobserved covariate could be explaining the results we see. The idea of the test is to introduce an unobserved covariate X that skews the probability of a unit being assigned to treatment or control, and then calculate by how much the probabilities would have to be skewed for the results observed to become statistically insignificant.

The results of the Rosenbaum sensitivity test are shown below in *Table 1*.

	Home to Neutral			Away to Neutral	
Gamma	Lower Bound	Upper Bound	Lower Bound		Upper Bound
1.0	0	0.0000	0		0.0000
1.2	0	0.0000	0		0.0000
1.4	0	0.0001	0		0.0000
1.6	0	0.0022	0		0.0005
1.8	0	0.0160	0		0.0065
2.0	0	0.0639	0		0.0369
2.2	0	0.1674	0		0.1227

Table 1. Results from the Rosenbaum sensitivity test for both control groups. The lower bound and upper bounds represent the p-values for the Wilcoxon Signed Rank test. P-values greater than 0.05, italicized, show that an unobserved covariate X that skews the probability of being assigned to treatment by 2.0 and 2.1 are sufficient to explain the observed effects, respectively. A full table is included in the appendix.

Thus, we see that for the home data, an unobserved covariate X that skews the probability of being assigned to treatment by a factor of 2.0 would be enough to explain the results obtained; for the away data, an unobserved covariate that skews the probability by a factor of 2.1 would be sufficient.

Neutral grounds have a positive impact on match outcomes by providing a balancing effect that reduces home advantages and therefore should be used more often by the Gaelic

Football and Hurling Associations

In this paper, we use genetic matching to obtain the causal effect of the influence of neutral grounds on match outcomes and find that they are very similar to those obtained using

coarsened genetic matching, even though the former method manages to retain a much larger percentage of the dataset for analysis and thus provides better generalizability and inferential validity. Thus, the genetic matching results reinforce what Kneafsey and Müller found in their original paper.

However, we also find that the results obtained are relatively sensitive, and thus not very robust. An unobserved covariate X would only need to skew the probability of being assigned to treatment by 2.0 and 2.1 for the home and away data, respectively, for the results observed to be statistically insignificant (and possibly explained by the unobserved covariate X). The fact that so many covariates were left uncontrolled for makes these results especially important, since one of these uncontrolled covariates could be driving the results we see for the effect of neutral grounds. For instance, the distance that a neutral ground is from the respective "home" of each team could lead to higher/lower attendance for one side, which could have an influence on match outcomes. Given that there are many possible covariates that were not used in the analysis, the sensitive results from the analysis cast some doubt on the robustness of the results obtained by Kneafsey and Müller.

In light of these findings, this paper asserts that neutral grounds have a positive influence on match outcomes, providing a balancing force that reduces the advantage the home team obtains by playing at home, and reduces the disadvantage that the away team faces by playing away. Because of this, we believe that neutral grounds are a good mechanism to improve the fairness of sporting competitions and should be used more often by the Gaelic football and hurling leagues to provide a fairer matching between teams without having to play two games separately at two different times. Additionally, using neutral grounds would represent a cost reduction for championships that must play two games (one at home and one away) to ensure fair results, not only in monetary terms but also psychological and physiological for the players involved. However, we believe that further research must be done controlling for more covariates to improve the robustness of the results, and thus provide more confidence about the exact effect of using neutral grounds. We also encourage more research to be done on international sporting leagues, which are more often forced to employ the home-away match paradigm to ensure a fair competition, as the time and cost savings of using neutral grounds would be much more consequential for these larger-scale events.

References

- Carroll, R., & Collins, K. (2012). An examination of home-field advantage in Gaelic football 2001–2010. In Proceedings of the 17th Annual Congress of the European College of Sport Science, 4–7 July, Bruges.
- Kneafsey, L., & Müller, S. (2018). Assessing the influence of neutral grounds on match outcomes. *International Journal Of Performance Analysis In Sport*, 18(6), 892–905. doi: 10.1080/24748668.2018.1525678
- Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, 1(1), 12–14.

Pollard, R., & Pollard, G. (2005). Home advantage in soccer: A review of its existence and causes. *International Journal of Soccer and Science*, 3(1), 28–44.

Pollard, R., Prieto, J., & Gómez, M. A. (2017). Global difference in home advantage by country, sport and sex. *International Journal of Performance Analysis in Sport*, 17(4), 586–599.

UEFA Champions League. (2019). Retrieved from <https://www.uefa.com/uefachampionsleague/> Appendix

Below is the balance obtained for the home and away datasets using genetic matching:

Home dataset:

```
Before Matching Minimum p.value: < 2.22e-16
Variable Name(s): competition_dummy_factor I(competition_dummy_factor * elo_diff_adjusted_log) I(sport * competition_dummy_factor) Number(s): 2 5 6

After Matching Minimum p.value: 0.52854
Variable Name(s): elo_diff_adjusted_log Number(s): 3
```

Away dataset:

```
Before Matching Minimum p.value: < 2.22e-16
Variable Name(s): competition_dummy_factor elo_diff_adjusted_log I(elo_diff_adjusted_log^2) I(competition_dummy_factor * elo_diff_adjusted_log) I(sport * competition_dummy_factor) Number(s): 2 3 4 5 6

After Matching Minimum p.value: 0.52351
Variable Name(s): elo_diff_adjusted_log Number(s): 3
```

Since both minimum p-values are well above 0.1, we can treat assignment to treatment (neutral grounds) as being as good as random.

Below are the full tables for the Rosenbaum sensitivity tests:

Home dataset:

Gamma	Lower bound	Upper bound
1.0	0	0.0000
1.1	0	0.0000
1.2	0	0.0000
1.3	0	0.0000
1.4	0	0.0001
1.5	0	0.0006
1.6	0	0.0022
1.7	0	0.0065
1.8	0	0.0160
1.9	0	0.0340
2.0	0	0.0639
2.1	0	0.1081
2.2	0	0.1674
2.3	0	0.2404
2.4	0	0.3239
2.5	0	0.4134
2.6	0	0.5038
2.7	0	0.5907
2.8	0	0.6705
2.9	0	0.7408
3.0	0	0.8006

Away dataset:

Gamma	Lower bound	Upper bound
1.0	0	0.0000
1.1	0	0.0000
1.2	0	0.0000
1.3	0	0.0000
1.4	0	0.0000
1.5	0	0.0001
1.6	0	0.0005
1.7	0	0.0021
1.8	0	0.0065
1.9	0	0.0167
2.0	0	0.0369
2.1	0	0.0712
2.2	0	0.1227
2.3	0	0.1922
2.4	0	0.2772
2.5	0	0.3729
2.6	0	0.4730
2.7	0	0.5710
2.8	0	0.6615
2.9	0	0.7410
3.0	0	0.8074

[1] The details of how Elo ratings are calculated can be seen in more detail in the original paper by Kneafsey and Müller (Kneafsey & Müller, 2018).

▼ Lending and Default Likelihood Prediction

https://prod-files-secure.s3.us-west-2.amazonaws.com/c4c0fb6a-3b8e-44a5-b1c4-2c5efdb2d812/83adc8f3-bba4-4b88-a68b-faf5a79a11c6/Lending_Club_EDA_Predictive.pdf

Finance

▼ Tax Compliance Market - Deep Dive

Below, I'll explore the opportunity and indirect tax compliance market:

Sales tax is very hard.

Sales tax is hard because the tax code is complex, unstandardized, and decentralized across 14,000+ jurisdictions in the US alone. Governments make businesses figure out how to comply with this complexity. Businesses are responsible for knowing where, when, and how much tax to collect; maintaining state registrations; and filing returns.

This is not just true in the US, it's a global problem.

There are new startups focused on solving it for other geographies like Europe (i.e., [Taxdoo](#)) and Latam (i.e., [Brinta](#)), while incumbents like Avalara are looking to expand their presence internationally.

Intro to Sales Tax.

This morning, when you bought a coffee from your favorite local cafe, you probably paid 6-10% more than the sticker price because of this pervasive and ubiquitous thing called sales tax. Sales tax is embedded in virtually all commerce in every country, state, and municipality around the world. It's a major source of revenue for state, local, and national governments.

Price preview ⓘ
Locations listed are based on your Tax registrations.

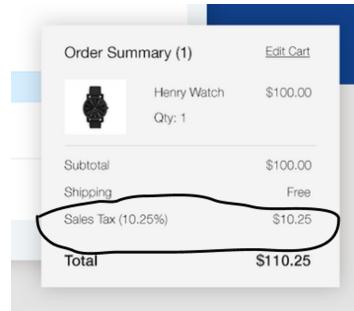
▼ California

Subtotal	\$10.00
Sales Tax (CA) 6%	\$0.60
Total	\$10.60

> New York

> Connecticut





The ancient Egyptians were the first civilization recorded to collect sales tax. Our modern sales tax system in the US traces its roots to the Great Depression when 25+ states established state-wide sales tax regimes to bolster revenue. Today, all but 5 states enforce sales tax. There is no escaping it if you consume or do business in our modern world. As Benjamin Franklin famously said:

"in this world nothing can be said to be certain, except death and taxes".

sam lessin 🎨 · [Follow](#)

wow i have never seen my state sales taxes broken out this way and now i am not happy! what are these taxes!? where did they come from!?

Subtotal	\$190.
CA STATE TAX (6.0%)	\$11.
SAN MATEO COUNTY DISTRICT TAX SP (2.125%)	\$4.
CA COUNTY TAX (0.25%)	\$0.
SAN MATEO CO LOCAL TAX SL (1.0%)	\$1.
Total	\$207.

4:39 PM · Aug 20, 2023

44 [Read 15 replies](#)

Why is Sales Tax so complex:

- Sales tax law is regulated by individual states, counties, and cities. There's no universal standard. The tax law and rates in Ohio are different than in California.
- The US has ~14k different taxing jurisdictions. In 2022, there were 6k+ tax law changes made within those jurisdictions. The rules are complex and can vary dramatically depending on the industry, products sold, where and how they're sold, etc.
- The tax code was written for a world in which commerce was primarily conducted in person. The internet changed all that. Today, every business is default national or global

from day one. Businesses of all sizes deal with sales tax on a scale that was previously faced by only the largest Fortune 500 enterprises.

- Historically, internet companies only had to collect sales tax in states where they had a physical presence, also known as "nexus" (retail stores, distribution centers, offices, etc.). Meaning most companies weren't collecting sales tax in the majority of states where they sold products. A 2018 ruling called Wayfair vs South Dakota changed everything, allowing states to enforce sales tax compliance on online purchases regardless of whether or not the company had a physical presence in the state. Today, 45 states and the District of Columbia enforce sales tax on internet companies.
- Marketplace facilitator laws are another recent regulatory change. These laws force marketplaces like Amazon, Walmart, Etsy, etc. to collect and file sales tax on behalf of their merchants. States started passing these laws in 2019. Currently, 30+ states and counting have passed some form of marketplace facilitator law.
- The growth in remote work has also increased the burden of tax compliance because hiring a contractor or FTE in a new state triggers an immediate tax compliance obligation, even if the company has no sales in that state.
- Merchants don't just sell physical products anymore. Now you can sell software, AI-generated art, e-books, data, subscription boxes, and whatever else entrepreneurs and creators make. Figuring out how your products are taxed is a challenge for modern companies that may not fit squarely into traditional boxes.

What customers expect

The last few years have seen tremendous innovation across B2B financial products. Today, you can set up payments (Stripe, Adyen); payroll (Gusto, Rippling, Justworks); banking (Mercury), corporate cards (Ramp, Brex, Expensify); hire globally (Deel, Remote), etc. — all in a matter of minutes. Hours at most. You don't need to be an expert. You simply need to click and let modern software do its job.

Finance Stack	Old generation	New Generation
Spend	 	   
Payroll	 	 
International EOR	 	  
Payments	   	 
Tax	  	?

Finance and accounting leaders have come to expect easy, intuitive, and elegant products, but that's not what they get when shopping for a sales tax solution. The fact that getting sales tax compliant is still so difficult is what initially sparked most new tax startups' curiosity.

Problems with existing software products

Today, you need to know a considerable amount about sales tax to successfully implement a sales tax product or pay someone who does. Avalara (the mid-market market leader) and Vertex (the market leader in the Fortune 500) sell complex enterprise tools meant to be used by in-house tax teams and/or set up by Big 4 accounting firms. Implementation can take months, hundreds of hours of engineering/accounting resources, and cost tens of thousands of dollars.

These are the most common frustrations...

- Implementation is time-consumptive, painful, and expensive
- Pricing is cost-prohibitive for growing companies
- UI is confusing
- Not developer friendly
- The product tax codes don't make sense for SaaS companies
- Mapping your products to the tax code is manual and confusing
- Customer support is terrible
- Missed or late tax filings
- Taxes being overpaid
- Inaccurate reporting data

Customer research

What I've seen from online reviews

Avalara fucked up our setup, which caused us to not collect the right amount of sales tax. The difference came out of our pocket.

CEO of an 8-figure ARR, Series A SaaS company

The product works if you get it set up right, but it just sucks. Their pricing is outrageous. They charge you for API calls that expire. So even if you don't use them all, you lose them going into the next billing period.

CEO of a venture-backed alcohol marketplace

It would have made our business unprofitable to comply with sales tax using Avalara so we decided to just risk getting audited until we have more revenue.

CFO of a bootstrapped \$7-figure digital products marketplace

From start to finish, it took a bunch of time to implement Avalara. Probably 8 weeks... 3 engineers, me, and my accountant.

CFO at a \$9-figure+ e-commerce marketplace

Because one of our products is taxed as Telecomm, we had to use two different APIs because the Telecomm product is different than the AvaTax product. Which made our engineers lives miserable for a couple months. We've actually still not finished getting it setup.

Controller at a ~\$40M ARR, Series B SaaS company.

I never know what I'm supposed to do with the information that Avalara gives me on a given state. Sales tax is risk management, and unless I pay them or an accounting firm 10k, I'm left shooting in the dark.

CFO of a 7-figure ARR venture-backed SaaS Company

We ended up with \$1M in liabilities plus penalties from different states because we started growing too fast. Our controller didn't know what to do. So we tried to go to Avalara to get it fixed, but they ghosted us on our first meeting then tried to charge us \$3k for something called a nexus study before they could give us a quote....

Founder of an 8-figure e-commerce holding company

Stripe Tax was too expensive and incomprehensive. Avalara was also very expensive, but at a certain business scale they're the only credible option.... We probably spent about 30-40 engineering hours on the Avalara implementation in total. Not including my time or my CFOs.

CEO of an 8-figure ARR Series A SaaS company

Our accounting firm told us to use Avalara.... They ended up registering us in too many states and then didn't do the registrations right in some states. So I had to go figure out how to manually log-in to each state's website and figure out what was going wrong.... Can't believe I spend \$50k a year on this.

CEO of an e-commerce marketplace

Implementation was a bitch. The docs were inaccurate. We had to get on multiple calls with their support engineers just to figure out what was going on.... The total setup ended up taking like 6 months.

Engineering Manager at \$80M+ Series D SaaS Company

I'd rip out Avalara tomorrow if a better product came long

CTO of a \$100M+ e-commerce brand

Avalara overpaid one state by \$100k for one of my clients. We couldn't even get ahold of their customer service to fix it.

Managing partner of a sales tax CPA / consulting firm

Market

This industry is the classic story of a dominant legacy incumbent starting to show important weaknesses that newer competitors are looking to exploit. Mainly, it is a red ocean market opportunity, with a go-to-market strategy that heavily relies on flawless execution.

The distinction, though, as early-stage investors, is that there are many startup competitors as well that have entered the market in the last 12 to 24 months.

What the market looks like today

Incumbents:

- Avalara: Comprehensive tax compliance solution with (almost) global coverage. Known for its strong branding but has faced customer frustrations. They were bought out by Vista for \$8.4B in October, 2022.
- Vertex & Sovos: Enterprise-grade solutions with deep tax expertise. Sovos is more prominent in Europe. Vertex is a publicly traded company with a \$3.5B market cap, and Sovos had been bought by Vista and more recently owned by HgCapital.
- TaxJar: User-friendly, e-commerce focused. Struggled with the SaaS market due to limited partnerships with billing platforms.
- Stripe: Attempted to dominate billing, payments and tax, but faced challenges in execution and integration post TaxJar acquisition.
- Thomson Reuters: a ~\$60B public company that also has a suite of legacy enterprise tax products.

Startups:

- Anrok (most similar to Taxwire): Focused on post PMF SaaS companies. They have the most developed product and business for SaaS sales tax of any startup to date. They

raised a \$20M Series A in 2022 at a \$100M valuation, led by Index Ventures and Sequoia Capital, and are rumored to have around \$5M in revenue.

- Zamp: Primarily focused on e-commerce brands. They've chosen to take a very services-heavy approach to what they're building.
- Numeral: A startup from the most recent Y Combinator batch. They're focused on handling filing & remittance for e-commerce companies. Biggest strength is that the founders have e-commerce backgrounds.

First, enterprise-graded companies are unlikely to switch providers. Most of the very large companies already have a solution (and those that don't may choose not because they do highly concentrated sales). The long term opportunity of targeting the younger SaaS market is to grow as many of these startups scale, but there's no immediate opportunity today to win big accounts.

Second, Avalara absolutely dominates the mid-market. Here, it is Avalara versus everyone else. Their core product was designed in the 2000s. The rest of it has been stitched together with the 18+ acquisitions they've made since. Internally, their tax filings are entirely manual. The product experience is clunky and overwhelming for non-experts. Their tax codes aren't configured for tech companies. Their international product is weak. For example, large enterprises like Mindbody (Avalara customer) are expanding internationally and choosing to go with alternative products for international tax compliance.

If you talk to Avalara's customers and partners, you'll discover that they've created a lot of frustration and animosity in the market. Since being acquired by Vista, customer service has declined, compliance mistakes increased, and pricing has gone up. Avalara's NPS is 13.

TaxJar (founded in 2013) was the closest thing to a real Avalara alternative, but Stripe has de-prioritized it since acquiring the company in 2021 (rumored at ~\$400M), laying off most of its team and halting GTM. Remaining efforts by Stripe in this space are focused on Stripe Tax.

Stripe Tax is not a true enterprise product. It can be good enough for small companies, but the fees, tax engine breadth, and lack of compliance features (filing, monitoring, registrations, etc.) prevent it from being a real solution for growing companies. Stripe Tax has access to Stripe data, and falls short in analyzing a company's broader context.

Global sales tax compliance is a growing market. States and countries aren't going to stop collecting taxes. Recent regulatory changes like Wayfair have increased compliance burdens. There are still many non-compliant companies that will need to get compliant. States are cracking down and the number of audits seem to be rising.

Concluding remarks

There are two main reasons why now may be good timing for a new player to enter the space: (1) market changes, and (2) increased regulations:

- Avalara has been acquired by PE, and their product isn't built for what many customers need today. They're also more focused on moving up-market (competing with Vertex for

Fortune 500 clients), cutting costs, and raising prices. They've created a lot of frustration among their customers and partners, and TaxJar is no longer a major player, so there's no clear challenger to Avalara.

- Regulatory changes like Wayfair vs South Dakota and marketplace facilitator laws coupled with increasing scrutiny from the states have made sales tax compliance a larger priority than ever.

However, there is no clear opportunity to leverage newer technologies (i.e., LLMs) to disrupt the industry. A few examples of suggested uses of LLMs:

- Facilitate a much faster onboarding process by understanding the business situation of customers using LLMs. During onboarding, users can explain their business in natural language, give a link they can scrape, so they can better automate the understanding and explaining of a company's taxability.
- Building a tax engine from scratch involves aggregating and interpreting large amounts of text and converting that into math. LLMs have already proven powerful for us in moving faster on understanding and interpreting the tax code.

The opportunity, as I see it, is to build best-in-class modern software with a great user experience that every company loves using. I've struggled to find a product approach that would afford a startup an unfair/intrinsic advantage in its go-to market, which is what we're looking for in software opportunities we back. Tax compliance is a low-margin-for-error problem space, offering no room for true innovation. At the end of the day, most startups can't control the regulations and how each State treats taxes, and companies are forced to play by their rules.

A big challenge is that the pull for tax compliance arises at specific growth milestones, so it is not a problem for most younger companies until it is.

▼ Rimba Diligence - Brainstorming

<https://rimba.cloud/>

~~Send follow up mike wang intro to watershed employees~~

~~Schedule mike mona ESG guy~~

Do manufacturers want to take this process in house?

Larger firms are more likely to attain certification to an environmental sustainability standard - 37 percent of companies with over 100 workers had such a certificate, while only 5 percent of firms with fewer than 5 workers had attained such a status.

Rimba is compliance software for manufacturers to streamline certification reporting for their annual audits. Their focus is on certifications like ISCC that focus on certifying sustainability across the supply chain. ISCC – as an example for this type of certificate – requires:

1. Traceability of sustainable products through the supply chain from the origin of the raw material to the final product.

2. Carbon emissions from the production of raw materials to the distribution of the final product.
3. Social and governance requirements suppliers must meet.

Broadly, compliance is a huge and growing issue in international trade. The requirements are numerous and vague which makes it difficult to implement and build a reporting system. 5 years ago, certifications were "nice to have". Today, regulatory pressures are making certification mandatory:

- EU = mandatory Renewable Energy Directive (RED) II since 2021
- Japan = mandatory through Japan's revised Feed-in-Tariff (FIT) since 2021
- USA = mandatory through Renewable Fuel Standard (RFS) since 2005

Their target market is companies who deal with feedstock in their supply chain, thus needing ISCC and RSPO certification. Their initial traction has been with biofuel manufacturers in Southeast Asia, but they are also having conversations with potential customers in the US and EU. Examples:

- Apical (biofuel trader & largest existing customer) needs ISCC to sell to Italy.
- Valero (petroleum manufacturing, \$55b market cap) needs ISCC Corsia to sell sustainable aviation fuel. CORSIA, the Carbon Offsetting and Reduction Scheme for International Aviation, is an emission mitigation instrument to address the increase in total CO2 emissions from aviation. 104 states (including all EU Member States), representing around 80% of international aviation, participate in CORSIA. Hence, the majority of airlines will be affected by the emission mitigation scheme.

ISCC and RSPO are the two top trusted certifications focused on biofuels and feedstock. RSPO is focused on palm oil and its derivatives, while ISCC has a broader certification scope for various feedstock.

TAM bottom up calculations

Knowns:

- RSPO has certified about 20% of the world's crude palm oil (CPO).
- ISCC has certified a total of 57,000 companies.

Assumptions:

- ISCC and RSPO have certified a similar % (so 20%) of their respective markets, as ISCC covers other feedstock (rice straw, wheat straw, rice husk, and corn stover, used cooking oil, tallow, animal fats, etc).

Number of companies globally that should be ISCC-certified: $285,000 = 57,000 * (100/20)$

Vanta pricing: For a company with a headcount of 200, the contract price ranges from \$16,100 to \$23,400. Companies with headcounts of 1,000 and over may expect to pay in the range of \$20,700 to \$92,200.

Assuming an annual contract value of \$25K to \$75K, it represents a market opportunity of \$7.1 to \$21.4 billion for Rimba to address.

Team

Distinct advantages to exploiting the Indonesian geography: Ryan was a Sr. auditor for certification frameworks like ISCC, RSPO, and RBO, and has built a network of renewable fuel suppliers and traders in Southeast Asia. Indonesia also represents 59% of the world's supply of palm oil.

It is currently 2 co-founders who previously worked together in this industry, and have signed their first five customers (all small to mid sized biofuel manufacturers). The founding team is not set. They are dating a third technical cofounder responsible for building the software: [Akshay Sharma](#) (who's already building for them), previously an ML engineer at [Klarity](#), a Series A AI documentation platform. They applied to Y Combinator this Monday, and I believe there may be an opportunity for us to invest before YC if they get accepted.

What does their product do?

Manufacturers have to go through thousands of docs (PDFs) for each transaction every week. Docs can be (1) certificates of weight (weights of raw materials they receive), (2) contracts with their suppliers, (3) delivery notes, (4) sustainability declarations from certified suppliers, (5) purchase orders, (5) certificate of analysis, (6) invoices – to name a few.

These manufacturers already track all of these for their manufacturing reports. The docs are in paper format, so they can scan them, and info is used for general ledger. If companies are not certified, they only care about the weights of each shipment, but if they are certified, they need to start tracking a lot more information such as:

- Traceability of raw materials
 - Where did they come from? Factory's location? Overlap with deforestation areas? Labor practices?
- Track real gas emissions
 - Calculate distance, type of trucks, how long of a journey, how much weight on trucks do they carry, depending on type of truck, type of fuel they use

Docs today are manually read, scanned, and added to multiple spreadsheets that track the aforementioned stuff. With all this data, companies create a mass balance spreadsheet that tracks all their inputs, and whether they are certified or not. This dictates how much % of their sales can go into markets that require certifications.

For example, if only 30% of their input materials are certified (in blended values) because they only manage to get 30% of their supply chain as certified. Then, only 30% of their sales can go into those markets that require certification (e.g., Europe and Japan). They can cross this limit, but it's a loan to future certified materials, so in aggregate, they can't cross this limit. If they do, their certification gets suspended.

The mass balance helps track this, how much certified material that comes in can't be more than the certified biofuel coming out.

Things Rimba is building to help streamline this:

1. Data Extraction and Reporting Automation:

- Process unstructured data e.g. scanned PDF, purchase order, weight certificate, shipping documents.
- Automates data extraction for compliance reporting; Intelligent Document Processing (IDP)

2. Material Tracking and Risk Management:

- Trace material flows and segregate certified from non-certified materials.
- Prevents double accounting and maintains integrity of supply chains.

3. Streamline Multiple Sustainability Certification Audit:

- more transparency + less manual data entry + less manual calculation on excel
- Automates cross-certification, reducing effort and duplication.

Competitive landscape

There are a number of successful companies that help demonstrate that there is a market here including Watershed, Greenly and Sustain Life.

Watershed is a software platform to help companies to zero carbon. Their mission is to build the tools that enable a business to:

- Measure their carbon emissions to know where they're at today.
- Plan out different ways to reduce the carbon emissions.
- Execute on that plan by buying clean power, by engaging their suppliers, by redesigning their product.
- Actually remove carbon from the atmosphere for whatever emissions remain (i.e., via their marketplace).
- Report on that progress externally so that investors, regulators, employees, customers, can hold them accountable (i.e., automate compliance for CSRD reporting) .

Podcast with CEO of Watershed (Scott Amyx):

Good news: There is starting to be some convergence and people are beginning to be in the space page about the rules of the road of what you count and how when tracking GHG. Bad news: The whole point of measurement is to enable the right action, and the status quo in carbon accounting maybe enables you to publish a really detailed PDF that has all the right acronyms on it, but does not enable companies to make the right choices. As an example, for most companies, most of their carbon emissions come from Scope3. The standard approach to calculating those emissions is to look at the dollars spent on a specific category and multiply it

by what the [EPA](#) said in 2010 is the carbon per dollar in that category. What Watershed wants companies to actually do is redesign their products, change materials and change suppliers, and the status quo doesn't enable that (i.e., just helps track but no actions). So Watershed wants to get the data to upgrade the quality of carbon accounting so it enables the right actions. So anyone can know the carbon footprint of a specific company, of a specific supplier, of a specific product, and you can make choices based on that.

In terms of regulations, the trend is for what's currently being enacted in the UK and Europe to become standard practice around the world. Large companies, both public and private, are expected to report on carbon accounting at least once a year with the same rigor with which they report their financial data.

This category of carbon accounting software has gained crazy momentum in the last few years. Notably, both Watershed and Greenly were founded in 2019 and have announced growth rounds in the last 60 days:

- [Watershed raised a \\$100M Series C led by Greenoaks, Kleiner Perkins, and Sequoia. Taylor Francis](#) is cofounder/CEO.
- [Greenly raised a \\$52M Series B led by Fidelity. Alexis Normand](#) is cofounder/CEO.

The POV of Watershed is that if companies have pressure to take responsibility for their entire carbon supply chain, they can bring capital, resources, and expertise to decarbonize some of these lesser developed countries, because they are part of that company's carbon supply chain. For example, when you look at a company like Apple, most of its carbon is coming from China and Taiwan from electronic suppliers (even though it appears on the US balance sheet).

This to me doesn't connect with stats like [RSPO having certified about 20% of the world's crude palm oil \(CPO\)](#). These producers of feedstocks like palm oil are based in the less developed countries that Scott (CEO) refers to in his podcast. 20% of crude palm oil having been certified already looks like incredible progress to me. Could it be that developed countries have this misconception that the biggest GHG emissions exist in less developed countries, when that might not be the case?

CSR (Corporate Sustainability Reporting Directive) is another type of certification that applies to large EU companies and public and private non-EU companies with EU business. It replaces the EU's legacy ESG reporting program—the NFRD—and raises the bar for breadth and robustness in sustainability reporting, covering categories beyond just carbon, including pollution, water, waste, and biodiversity.

They target companies in the EU and the US who are mostly concerned about carbon emissions (doesn't matter if they are manufacturing, banks, or software companies). They are concerned about the mandatory disclosure they have to make about carbon emissions (i.e., CSR). This is why companies like Greenly and Watershed have taken off.

Watershed also lets companies easily repurpose their CSR data to fulfill other reporting frameworks, like [CDP](#) or [ISSB](#). As an aside, only people who care about these reports are

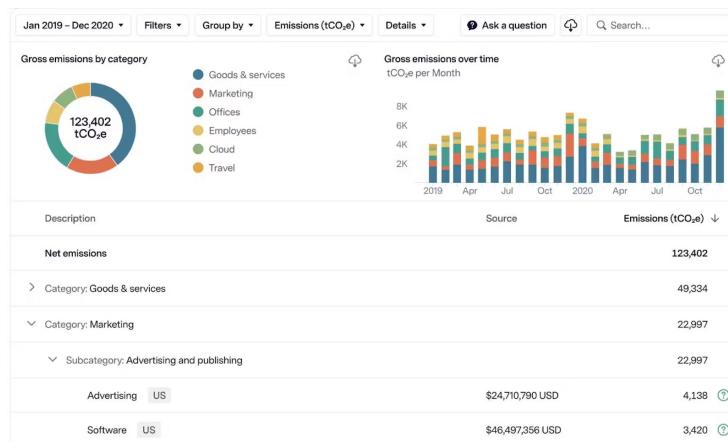
investors focused on ESG.

Watershed's vision is to become the one-stop shop for corporate sustainability teams. They offer everything these teams need. It's different to an approach like Rimba's that's going deeper into the end-to-end compliance requirements for a specific vertical (manufacturers) versus targeting a specific team (corporate sustainability) across many verticals. This shows in their product suite:

- Corporate: reporting; e.g., CSRD.
- Financial services: data to analyze portfolio's emissions (ESG stuff)
- Supply chain visibility: gives corporations visibility over which suppliers are generating most carbon emissions (scope3 emissions i.e., GSG coming from suppliers indirectly), so they can grill them down and try cutting down their own carbon footprint.
- Consumer goods: help track carbon emissions for consumer good brands that care about consumer sentiment (i.e., make more money if they can show they are sustainable).
- Carbon Offset Marketplace: companies can purchase carbon offsets commensurate with their unavoidable emissions. The market for carbon offsets is expected to grow from \$2 billion in 2020 to around \$250 billion by 2050.

Watershed claims to streamline emission measurement using customer data to estimate a company's Scope 1 (owned assets emissions), Scope 2 (utility purchases emissions), and Scope 3 emissions (3rd party emissions). For example, Watershed will use an Excel spreadsheet or a purchase order PDF to estimate emissions. Watershed also works with a company's suppliers to accurately assess a company's Scope 3 emissions. In November 2022, Watershed announced a supply-chain emissions tracking module to better address Scope 3 emissions with a "combination of public and private data from suppliers to estimate the emissions footprint of the larger supply chain."

A dated view of Watershed's product that is no longer on its website shows an expense item-level drilldown to illustrate the exact emissions coming from particular expense areas (e.g. advertising, software, etc.).



Watershed's takeaway: The majority of Watershed's customers are technology companies. If climate reporting becomes mandatory, the market size will increase dramatically as corporations scramble to become compliant. You could potentially compare Watershed to ERP platforms like Oracle or SAP, or vertically-focused ERPs, like Veeva or Toast. If Watershed is successful the company would be considered a core system of record for managing carbon accounting. These types of ERPs can trade anywhere from 2-11x LTM revenue. Though in this podcast, Watershed CEO does describe their focus beyond carbon accounting, "we're not in this to play the carbon accounting game, we're in this to be a platform that enables decarbonization and companies at the front of climate change to make different choices".

Why is Rimba different?

Rimba is focused on a different niche. They noticed that aside from carbon emissions, regulators are also starting to ask for disclosure of where the raw materials are coming from, social and governance aspects of these suppliers, and carbon emissions from the transportation that is involved in the supply chain until the consumer (upstream data). These data are captured in other certifications like ISCC.

Watershed and Greenly haven't achieved any meaningful penetration more upstream in the supply chain (producers, traders, farmers). Most of their clients are startups, financial institutions, and public companies. When Rimba talks to potential customers like Valero (petroleum manufacturing, \$55b market cap), IAG (International Airlines Group), or other manufacturers, they never mention Watershed as someone that is helping them.

Where Rimba comes in is that all these companies fail to understand how to navigate legacy industries like agriculture, manufacturing, and mining, which are all growing massively in Asia and LatAm supplying these products to the EU and the US. Most companies outside of the US and the EU probably never heard of CSRD certification or carbon emissions accounting. All they care about is how they can export products overseas, and for that they need to be certified with other kinds of certificates like ISCC, RSPO, and RBO.

Well, this still begs the questions.

1. Is it the logical next step for companies like Watershed to start adding other certificates to their platform after they win over CSRD compliance? If yes, which certificates are they likely to go after next? If the market opportunity for certificates like ISCC is obviously large, what are the reasons for why they wouldn't choose to focus on that? Are there key differences between these markets? Is it a wholly separate problem/focus?
2. Why would Watershed's lack of experience in legacy markets affect their chances of winning? What if the only important segment that needs to be won are large corporations serving the end consumers? One could say that if you win over this segment, you can help them force similar measures across all their suppliers upstream.

Per their CSRD website, Watershed's likely to focus on the certifications next:

How does CSRD compare to other ESG disclosures?



Interesting potential customer use case:

Upstream data is unstructured (many suppliers are not sophisticated even in the US). Tim talked to the sustainability officer of a company called Highland Pellets and they sell biomass to the EU from the US. Highland sources wood chips and shingles from lumber mills and privately owned forests in the US. Many lumber mills and forest owners do not want to disclose anything. It is a constant push and pull, but Highland must be certified by certification: Sustainable Biomass Program / ISCC, and they need to track every square foot of where the trees come from. The requirements a few years ago were not very stringent, but it has gotten harder.

- If it is a certified supplier they would need less verification of supplier data. They would focus on delivery notes, sustainability certificate, and sustainability declaration.
- If the supplier is not yet certified, the company needs to verify the data themselves. (Like doing internal audit of the supplier)

Highland's an interesting example as it shows how manufacturers struggle to get the necessary data from their suppliers. In Highland's case the requirement is to trace where each tree comes from in which part of the forest. So Rimba has been discussing a solution by building a phone app for suppliers to take pictures of the trees before they are cut down, and Rimba will connect the app to use the photo and google maps to pinpoint the location. Right now Highland doesn't have a solution for this yet.

CDP/SDDB/CSRD's main audience is investors, these frameworks are meant to help public market investors determine which companies are better from a carbon accounting perspective.

ISCC on the other hand is more focused on operational data, and has a direct impact on a company's ability to operate and do business. In order, it focuses on:

1. Traceability of raw materials

2. Deforestation, and then,
3. Carbon emissions

CarbonChain is a competitor that's more similar to Rimba: Founded in 2019, CarbonChain is a carbon accounting platform specifically targeted at manufacturers, commodity traders, and their banks. Companies tied to physical supply chains have a more complex emissions footprint, and can also access unique sustainable financing if they can demonstrate emissions measurement and reduction. Customers include Lloyd's, Concord, and Gunvor.

CarbonChain Comply offers a comprehensive solution for businesses within metals and energy supply chains who are grappling with monumental regulatory shifts on carbon emissions and rising stakeholder demand for transparent and accurate disclosure. The new one-stop shop for carbon reporting empowers sustainability leaders, compliance managers and procurement directors to streamline and satisfy every key disclosure request, no matter the format or framework. It raised a \$10M Series A led by Union Square Ventures on April 10, 2024.

They are starting to talk about this other EU requirement called CBAM targeted at importers of aluminum, steel, iron, fertilizers, electrical energy, hydrogen or cement into the EU.

Interesting case study: <https://www.carbonchain.com/case-study/concord>

Detailed description of ISCC requirements (pdf p19 to p24):

The ISCC certification system covers three categories of requirements:

- > Sustainability requirements for agricultural and forest biomass, waste and residues, feedstocks for the production of renewable fuels of nonbiological origin and recycled carbon fuels and low iLUC risk feedstocks
- > Requirements for traceability and the chain of custody
- > Requirements for greenhouse gas emission savings and the calculation methodology

Sustainability requirements

Farms and plantations that produce sustainable agricultural biomass must comply with the sustainability requirements as stated in ISCC EU System Documents 202-1 "Agricultural Biomass ISCC Principle 1" and 202-2 "Agricultural Biomass ISCC Principles 2-6" ". The six ISCC Principles cover the following topics:

- Principle 1: Protection of land with high biodiversity value or high carbon stock
- Principle 2: Environmentally responsible production to protect soil, water and air
- Principle 3: Safe working conditions
- Principle 4: Compliance with human and labour rights and responsible community relations
- Principle 5: Compliance with land rights, laws and international treaties
- Principle 6: Good management practices and continuous improvement

If biofuels, bioliquids and biomass fuels are produced from forest biomass the risk of using forest biomass from unsustainable production has to be minimised, and the biomass has to meet certain land-use, land-use change and forestry (LULUCF) criteria. The use of forest biomass from sustainable production shall be ensured through monitoring and enforcement systems on national level or through requirements for the management system of the harvesting operation (Art. 29 (6) of the RED II). The six ISCC Principles cover the following topics:

Principle 1: Sustainability requirements for the production of forest biomass

Principle 2: Environmentally responsible production to protect soil, water and air

Principle 3: Safe working conditions

Principle 4: Compliance with human and labor rights and responsible community relations

Principle 5: Compliance with land rights, laws and international treaties

Principle 6: Good management practices and continuous improvement

Traceability and Chain of Custody

Traceability and chain of custody ensure that two basic requirements are fulfilled:

1. The possibility of tracing sustainable products back and forth through the supply chain from the origin of the raw material to the final product
2. The possibility of assigning product specific information to consignments (batches) of sustainable materials and products

Traceability describes the ability to identify and trace the origin, processing history and distribution of materials and products through supply chains. Chain of custody describes the monitoring of input and output of sustainable materials and related information and documentation requirements. The following properties of sustainable material (so-called sustainability characteristics) are the minimum information required under ISCC:

- > Type of raw material (e.g. rapeseed, sunflower, used cooking oil, etc.)
- > Country of origin of the raw material
- > Certification scope of raw material (e.g. the raw material is either certified according to the sustainability criteria of Art. 29 (2) – (7) of the RED II, or the raw material meets the RED II definition of waste or residue)
- > Information on GHG emissions (mandatory under ISCC EU, voluntary under ISCC PLUS)
- > Claim "ISCC compliant" or "EU RED compliant" (if applicable)

Greenhouse Gas Emissions

ISCC provides a methodology for calculating greenhouse gas (GHG) emissions for all elements of the supply chain and for determining greenhouse gas emissions savings. This can be applied to supply chains in all markets. Specific requirements apply for biofuels, bioliquids, biomass fuels, recycled carbon fuels and renewable liquid and gaseous transport fuels of nonbiological origin that are brought into the markets of the European Union. The economic operators bringing sustainable fuels into the market (i.e. the fuel supplier) must prove that the

fuels where produced sustainably and that greenhouse gas emissions savings are sufficient according to the RED II.

The requirements for GHG emissions apply to all relevant supply chain elements from the production of raw materials to the distribution of the final product, including cultivation, collection and conversion processes, as well as the transport and distribution of intermediate and final products. Three different options are available to provide information on GHG emissions:

1. Use of total default values: Default values are raw material and process specific and are provided in the RED II for different types of biofuels, bioliquids and biomass fuels.
2. Use of disaggregated default values. So-called disaggregated default values are available in the RED II for the cultivation/ production of biomass, processing, and transport and distribution. Disaggregated default values allow the use of a combination of default and actual values.
3. To determine GHG emissions from cultivation the use of typical values that represent the average value in a specific area is also possible. Those typical values are often referred to as NUTS210 values. Those NUTS 2 or typical values have to be reported to the European Commission by Member States or third countries and can be used if the Commission recognizes them to be accurate . Use of actual values: Individually calculated values must be calculated based on the methodology according to the RED II.

All requirements to apply, calculate and verify the greenhouse gas emissions and emission savings are specified in ISCC EU System Document 205 "Greenhouse Gas Emissions".

Participants in the Certification System

All economic operators that handle sustainable material (e.g. produce or generate, collect, process, store or trade) have to be covered by an ISCC certification. In the ISCC system the term 'economic operator' refers to a specific site (spatial entity) of a company where sustainable material is handled. A company may have more than one operating sites. If this is the case, every single operating site handling sustainable material has to be covered by its own certification.

For biofuels, bioliquids and biomass fuels produced from agricultural or forest biomass (including agricultural, aquaculture, fisheries and forestry residues) the supply chain starts at the farm/plantation or forest management unit (see Figure 2). The first gathering point is the first element that must be individually certified. Farms/plantations and forest management units can be covered under the certificate of the first gathering point but may also receive an individual or group certification.

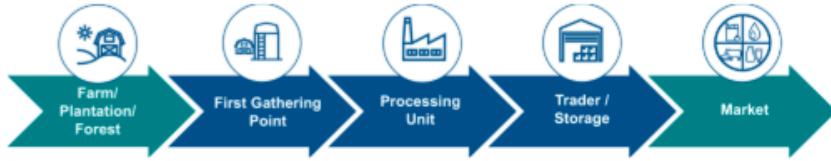


Figure 2: Example of simplified supply chain for agricultural feedstocks/agricultural crop residues and forestry feedstocks/forestry residues

Biofuels, bioliquids and biomass fuels can also be produced from bio-based waste and processing residues. Recycled carbon fuels are produced from non-renewable liquid or solid waste. For renewable liquid and gaseous transport fuels of non-biological origin (RFNBO) it is the process energy that provides the energy input for the final fuel (e.g. electricity generated from wind, solar, aerothermal, geothermal or water). This process energy has to be derived from renewable sources other than biomass.

In these cases, the supply chain starts at the point of origin (see Figure 3). The collecting point is the first element that must be individually certified. Points of origin can be covered under the certificate of the collecting point but may also receive an individual or group certification.

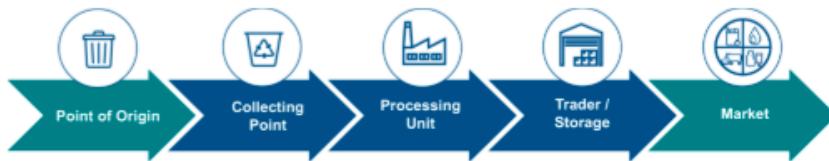


Figure 3: Example of simplified supply chain for waste, processing residues and renewable non-bio feedstocks

For all elements of the supply chain after the first gathering point or collecting point the certification requirements are the same regardless of the kind of raw material that enters the supply chain.

A valid ISCC certification is a prerequisite for any element of the supply chain that will handle sustainable material. The only exception to do this is for first gathering points and collecting points that may receive sustainable material up to three months prior to the start of the certificate's validity. A detailed description of all relevant supply chain elements, the respective certification requirements and the requirements for handling sustainable material are stated in ISCC EU System Document 203 "Traceability and Chain of Custody".

More random notes all over the place below

Their current workflow: Drive folder with all the pdfs per type, spreadsheet per type, mass balance report

Rimba is compliance software for manufacturers to streamline certification reporting for annual audits. They're developing an AI enabled platform that uses LLMs and OCR to automatically extract key information from unstructured data sources (e.g. PDFs, scanned delivery notes,

etc.). This addresses the challenge of manually handling thousands of such documents weekly, reducing human error leading to the risk of certification suspension.

The platform ensures accurate tracing of raw material origins and GHG (greenhouse gas emissions) calculation from transportations involved, which is crucial for meeting certification requirements. Based on customer feedback, this saves their customers from having to hire consultants, more employees to maintain certification, and reduce risk of suspension.

Their long-term goal is to expand into certifications in other industries such as textiles, plastics, food. They'll first expand their product capabilities to integrate with various ERP softwares and – applying AI to cut down on manual review and calculations currently being done by consultants.

By accessing all data that currently lives on PDFs, they'll focus on the following key features:

- Dashboard for visibility
- Trace material flows and segregate certified from non-certified materials.
- Prevents double accounting and maintains integrity of supply chains.
- Algorithm to match volume and source of purchased materials and sold materials
- Automates multiple certifications, reducing effort and duplication.

Rimba is compliance software for manufacturers. They streamline and automate how manufacturers follow mandatory sustainability certifications like ISCC, RSPO, RSB, starting with the renewable fuel market.

The market opportunity has grown and became more relevant in the last few years. The reason is that these sustainability certifications have become mandatory and require very stringent guidelines. Companies that don't become certified can't do business in the US, Europe, and Japan. The process for becoming certified and keeping the certification via annual audits lives in spreadsheets, manual work, and only happens thanks to expensive consultants. Needless to say, this is an expensive, time consuming, manual process. New technologies like OCR and LLMs allow for useful automations that can truly make a dent here, so there was never better timing to build a software company that helps automate and manage compliance for this market.

It is currently 2 co-founders who previously worked together in this industry, and have signed their first five customers (all small to mid sized biofuel manufacturers). The founding team is not set. They are still looking for a third technical cofounder responsible for building the software. They are actively dating CTOs and seem advanced with this one engineer called Akshay Sharma (who's already building for them), previously an ML engineer at Klarity, an AI documentation platform. They are applying to Y Combinator by Monday, and I believe there may be an opportunity to invest before YC if they get accepted.

What does their product do?

Manufacturers have to go through thousands of docs (PDFs) every week. Docs are either (1) certificates of weight (weights of raw materials they receive), (2) contracts with their suppliers,

and (3) delivery notes.

These manufacturers already track all of these for their manufacturing reports. The docs are in paper format, so they can scan them, and info is used for general ledger. If companies are not certified, they only care about the weights of each shipment, but if they are certified, they need to start tracking a lot more information such as:

- Traceability of raw materials
 - Where did they come from? Factory's location? Overlap with deforestation areas? Labor practices?
- Track real gas emissions
 - Calculate distance, type of trucks, how long of a journey, how much weight on trucks do they carry, depending on type of truck, type of fuel they use

Docs today are manually read, scanned, and added to multiple spreadsheets that track the aforementioned stuff. With all this data, companies create a mass balance spreadsheet that tracks all their inputs, and whether they are certified or not. This dictates how much % of their sales can go into markets that require certifications.

For example, if only 30% of their input materials are certified (in blended values) because they only manage to get 30% of their supply chain as certified. Then, only 30% of their sales can go into those markets that require certification (e.g., Europe and Japan). They can cross this limit, but it's a loan to future certified materials, so in aggregate, they can't cross this limit. If they do, their certification gets suspended.

The mass balance helps track this, how much certified material that comes in can't be more than the certified biofuel coming out.

Things Rimba is building to help streamline this:

1. Data Extraction and Reporting Automation:
 - Process unstructured data e.g. scanned PDF, purchase order, weight certificate, shipping documents.
 - Automates data extraction for compliance reporting; Intelligent Document Processing (IDP)
2. Material Tracking and Risk Management:
 - Trace material flows and segregate certified from non-certified materials.
 - Prevents double accounting and maintains integrity of supply chains.
3. Streamline Multiple Sustainability Certification Audit:
 - more transparency + less manual data entry + less manual calculation on excel
 - Automates cross-certification, reducing effort and duplication.

Some context on how the market works:

Independent auditors run third party verification services that audit whether manufacturers are following certification guidelines.

Consultants help train and prepare manufacturers for these audits. In practice, manufacturers only learn 20-30% of what consultants tell them in Year 1, so they continue paying for their services annually to make sure they don't get suspended in the annual audits.

Manufacturers also hire an in-house team (typically under the legal department) that make sure all the required data is gathered.

ISCC and RSPO are focused on biofuels and feedstock. 50K companies have been certified, a 25% of those that need it, so the total number of companies here is 180-200K.

Knowns:

- Some certifications are bullshit (self reported/voluntary guidelines), but others are real (third party verification via regular audits) and have real economic consequences (not allowed to sell in various markets).
- Acquiring and maintaining these real certifications is not an easy task. It's manual, time consuming, and expensive. Currently, consultancy services like Control Union are the biggest players in helping manufacturers navigate the certification requirements (5000 employees, \$600M annual revenues).

Unknowns:

- Unclear how much money manufacturers spend on consultancy services + in house employees to navigate certification processes.
- TIC companies are involved with this, but unclear how relevant is certification for the largest companies in this space.

Testing, Inspection, and Certification (TIC) market.

- Certification is part of testing, inspection and certification and the provision by an independent body of written assurance (a certificate) that the product, service or system in question meets specific requirements. It is the formal attestation or confirmation of certain characteristics of an object, person, or organization. This confirmation is often, but not always, provided by some form of external review, education, assessment, or audit.

The Global Testing, Inspection, And Certification (TIC) Market is projected to grow at a CAGR of 5.5% from 2022 to 2030. the market was valued at USD 233.69 Billion in 2021 and is expected to reach USD 389.79 Billion by 2030.

- SGS S.A. (\$7B annual revenues, public)
- Bureau Veritas S.A. (\$6B annual revenues, public)
- Intertek Group Plc. (\$3.5B annual revenues, public)

Testing, Inspection, And Certification (TIC) Market industry is experiencing unprecedented growth, driven by regulatory compliance, outsourcing trends, environmental awareness, and technological advancements

- Rising Regulatory Compliance: Governments across the world are enforcing stringent regulations to ensure the safety and quality of consumer products, compelling manufacturers to engage independent TIC services.
- Outsourcing Trend: Manufacturers are increasingly outsourcing TIC services to specialized providers, aiming to meet compliance standards while streamlining operations and enhancing profitability.
- Environmental Awareness: Growing environmental consciousness has led to a surge in environmental testing and certification, driven by both governmental mandates and consumer demand.
- Technological Advancements: Continued innovation in testing, inspection, and certification technologies will enhance the industry's efficiency and capacity to address emerging challenges.

Another relevant market: revenue attributed to environmental certificates is expected to increase from \$37.8 billion in 2023 to \$102.3 billion in 2032, at a compound annual growth rate (CAGR) of 11.7%.

Bottom up TAM analysis on SOC2 compliance annual spend:

<https://start.askwonder.com/insights/market-size-compliance-audit-software-uogrw344c>

*Therefore, we can assume that the total estimated SOC 2 spend in the U.S. every year would be $(75\% * 9,000) * 60,000 = 405,000,000$.*

- Total number of SaaS companies: 9000 [USA]
- Avg annual audit spend: \$20k to \$100K
- 75% of SaaS companies get SOC certification

Similar numbers for ISCC alone:

- Total number of ISCC certified manufacturers: 50,000 [Global]
- Avg annual audit spend: unknown; Tim's mining company \$20K.

Customers

Focused on the Indonesian market for palm oil; Indonesia produces 44.5 million metric tons of palm oil, or roughly 59 percent of the world's supply. It's in so many everyday products around the world—shampoo, deodorant, cereal, dishwashing detergent—that much of the world could scarcely function without it. Yet the massive industry is still lacking in many basic regulations. The high-yield nature of palm oil, which yields almost ten times as much product per hectare as soybean oil, makes it a lucrative prospect for corporations. And its expected market growth could nearly double to £87 billion (over \$105 billion in US dollars) by 2026—what Bintang's team calls “an agricultural gold rush.”

Largest customer: <https://www.apicalgroup.com/sustainability/progress-dashboard/>

- <https://rspo.org/members/2-0235-11-000-00/>

Klarity

A disproportionate percentage of highly skilled, highly paid workers waste time reading similar documents over and over again.

Klarity's platform was built over the course of four years and leverages cutting-edge Natural Language Processing and Computer Vision to understand documents as well — or better — than human reviewers. The result: The most accurate document intelligence platform available today.

Market size

Environmental certificates is a growing market. Seemingly because the carbon and energy attribute certificates can be traded in carbon markets, and a lot of money can be made that way. It's still unclear to me how third-party verified certificates like ISCC for biofuels or Energy Star for energy efficiencies are connected to this. My guess is that if I am a biofuel producer, and now I get 40% of my production to be certified materials, perhaps I can create credentials from that that I can later sell in the markets?

Unclear the market size breakdown. How much of it is related to financial markets stuff like carbon bonds, certificates that are sold or bought by other companies to offset carbon footprint versus the actual cost of managing certification compliance. I assume that number is much smaller.

Different kinds of certifications:

- Self reported or voluntary guidelines
- Third party verification and mandatory

like ISCC, RSPO, and RSB in the context of BioFuels

His plan:

- Technical cofounder
- 10 companies under their belt

They have 5. Get another 5.

RIMBA is a compliance automation platform for manufacturing companies globally to meet industry-standard sustainability certifications.

I am looking to understand what is everything that needs to be tracked. Why not just narrow down their focus to ISCC to start? Each certification seems to be a pretty unique process (different to what he somewhat suggested in the first call that they are all kind of the same). I want to understand in detail how can software truly move the needle for ISCC (specifically one example).

1. Extracting data from invoices.
2. Who does traceability? Are companies good at this with or without software? Startups seemed to tried doing traceability but not from compliance angle. Thoughts?

3. Who tracks traceability, and who tracks greenhouse gas emission savings? How do companies get to these numbers? How do they create their methodology?

Isn't \$200k a bit too low for this to be a big problem?

ISCC requirements:

3.2 Certification Requirements

The ISCC certification system covers three categories of requirements:

- > Sustainability requirements for agricultural and forest biomass, waste and residues, feedstocks for the production of renewable fuels of non-biological origin and recycled carbon fuels and low iLUC risk feedstocks
- > Requirements for traceability and the chain of custody
- > Requirements for greenhouse gas emission savings and the calculation methodology

These certification requirements and the relevant processes are outlined in detail in the respective ISCC System Documents. No changes to the requirements as set in the RED II are permitted. The implementation of requirements that go beyond the requirements of the respective version of the RED is possible under ISCC.

Product traceability in supply chain is the process of tracking a product's origin and journey through the supply chain, from manufacturing to end-use. It involves logging manufacturing information in a database and following products at every step.

3.2.2 Traceability and Chain of Custody

Traceability and chain of custody ensure that two basic requirements are fulfilled:

- 1 The possibility of tracing sustainable products back and forth through the supply chain from the origin of the raw material to the final product
- 10 The possibility of assigning product specific information to consignments (batches) of sustainable materials and products

Traceability describes the ability to identify and trace the origin, processing history and distribution of materials and products through supply chains. Chain of custody describes the monitoring of input and output of sustainable materials and related information and documentation requirements. The following properties of sustainable material (so-called sustainability characteristics) are the minimum information required under ISCC:

- > Type of raw material (e.g. rapeseed, sunflower, used cooking oil, etc.)
- > Country of origin of the raw material
- > Certification scope of raw material (e.g. the raw material is either certified according to the sustainability criteria of Art. 29 (2) – (7) of the RED II, or the raw material meets the RED II definition of waste or residue)
- > Information on GHG emissions (mandatory under ISCC EU, voluntary under ISCC PLUS)
- > Claim "ISCC compliant" or "EU RED compliant" (if applicable)

Lack of product expertise. As it is right now, I worry they don't have the experience or talent to build a great modern product to serve how this industry manages compliance. They claim to be Vanta for this sustainability certification niche, but I worry in practice they won't get close to building a similar quality product to Vanta's. However, I do think there is a good founder market fit with the CEO and this space, and if somehow paired with the right CTO, this could become a killer team.

- Beyond the idea. We want to see a first-gen product that's delivering value.
- Product-centric opportunities, e.g. fundamentally superior product.
- PMF focus. Startup with a maniacal focus on validating product-market fit.
- Builders. The product is a core competency of the team, not an afterthought.
- True entrepreneurs. Not in it for the immediate gratification, glory, lifestyle.
- **Deep insight. Founders understand the problem better than most people in the world.**
- Blue ocean strategy. Product creates new market space rather than competing head-to-head.

- Wedgeable. Nearterm path to dominating an initial, narrow wedge of a larger market.
- Interesting IP created by solving very hard problems worth solving.
- Venture scale. The business should be raising VC to capitalize on its opportunities.
- **Scalable business model. If not initially, then very possible in the intermediate term.**
- **Durable business model. The company can survive/thrive outside of trends or economic cycles.**
- **Execution. There is evidence that the team can do the hard work.**
- Pre-A. We will work with founders to get them into a nice series A.

Market

- Does this match our stated areas of interest?
 - If not, does the opportunity represent a new potential area of interest that we should further research?

It has the potential to be interesting. Staying compliant with many of these certifications looks like a cumbersome, manual process where software has very little penetration. If built the right way, the opportunity of building a Vanta for this type of compliance seems appealing to me.

- Do we intuitively understand and agree with the problem?

Obtaining and managing a sustainability certification is a manual, time consuming process. If it costs \$200K a year to manage compliance (auditors, consultants, skilled labor, and certification) for a \$20M revenue company, it looks like a relatively high cost.

- Do we agree with the company's perspective on the market?
- Is this a blue ocean or red ocean market?
- How big is the market, both immediate and potential?

No numbers so far but relatively large. There are a lot of manufacturers that need to be certified, and for that have to manage their business.

- What are examples of successful companies that demonstrate that there is a market here?

There are no successful large scale product startups in this space.

This one seems to have failed: <https://jobs.techstars.com/companies/bioledger>. They were trying to build a single source of truth for all transfers in the lifecycle of a biofuel. Not sure what they were trying to do or who this was for.

Product traceability in supply chain is the process of tracking a product's origin and journey through the supply chain, from manufacturing to end-use. It involves logging manufacturing information in a database and following products at every step.

<https://www.veriflux.io/>. This other competitor also emphasizes the traceability aspect of this. RIMBA didn't discuss this as much. Why?

- If not based in the US, are there distinct advantages to exploit in that geography?
- Is there a wedge approach to the market?
- Is the business model scalable to the degree that warrants venture capital?
- Is the business model durable enough to survive surprising dips in a market cycle or unexpected tailwinds?

Team (Founders)

1. Empathy: This entails a deep understanding of the customer's needs and the broader market pain points. It's not about superficial knowledge, but rather a profound comprehension that fuels the team's instincts and drives their decision-making process. This empathy allows them to connect with their customer base on a more intimate level, creating products and services that truly resonate.

2. Expertise: This isn't just about having domain experience—it's about possessing a level of understanding so deep that it sets the team apart from others. Their expertise should enable them to navigate complex industry challenges and uncover unique insights that can be turned into actionable strategies.

3. Execution: This is all about the nitty-gritty of operations. The team's ability to execute plans efficiently and resourcefully is key. It's not just about having a great idea—it's about being able to bring that idea to life in the most effective way possible.

4. Experimental: This quality is characterized by a willingness to build, test, learn, and iterate. It's about having an open mindset and the flexibility to adapt and evolve. A team with an experimental approach is unafraid of making mistakes—they see them as learning opportunities that can lead to improvement and innovation.

- Are the founders good at communicating the problem?
 - Did they teach us anything new?
- Do they understand this problem better than most people in the world?
 - What gives that insight or experience?
- Is this the right team to capitalize on the opportunity? Even if they have the right insight, make sure that this is the team to do it.
- Are the founders hands-on builders?
 - What have they built in the past?
- Who is responsible for difficult, technical work on the team?
- Would we describe this team as true entrepreneurs?
 - True entrepreneurs are resourceful and will survive the low points and exhausting grind of building a business. They are not in it for the immediate gratification, glory, lifestyle, or treat it as a stepping stone in their careers.

Product

- Is the company working on directly addressing the “problem,” or are they working adjacent to it?
- Is it clear who is the intended audience for this product?
- Is the product built and launched?
- What makes the product fundamentally superior or different from other options?
- Does the product deliver value to users/customers in its current state?
 - What are the top 2 reasons that users/customers get value from the product?
 - What are the top 2 improvements that users/customers would benefit from?
- Is there interesting technology or IP being built here?
- What are the associated costs to this product? Consider the cost of building and distributing the product.

Traction (Momentum)

- Do the founders believe they have validated a minimal degree of PMF?
- How do the founders view the process for figuring out PMF?
- How has this team shown that they can execute on their potential?
- What are recent inflection points that make this a good time to invest in the company?
- What are the primary drivers to their growth?

Opportunity

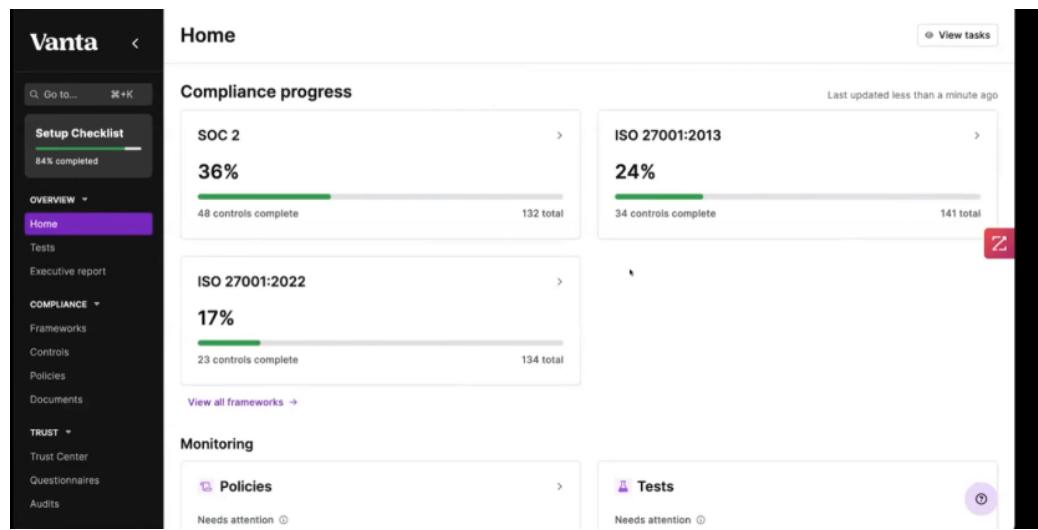
- Once they raise money, what is unlocked for the company moving forward? How does this accelerate them?
- Who are the Series A investors that we think would be interested in this company, if they hit the right milestones?

What big questions have been answered so far? Derisk. Validation/validation

Before solutions like Vanta, obtaining SOC2 certification was generally more manual and time-consuming. Here's a breakdown of how it typically went:

- Understanding the Requirements: Companies had to deeply understand the SOC2 criteria, which involve detailed controls on security, availability, processing integrity, confidentiality, and privacy. This meant a lot of research and possibly hiring consultants to ensure full comprehension.
- Manual Documentation: Businesses needed to manually document all processes and controls relevant to the SOC2 standards. This often required extensive efforts from internal teams or expensive external consultants to map out processes, write policies, and prepare the required documentation.

- Implementing Controls: Implementing the necessary controls without a guiding software meant relying on internal knowledge and often bespoke solutions. This step could include configuring systems to meet security standards, establishing new protocols for data handling, and training employees on compliance-related issues.
- Evidence Gathering: Companies had to manually collect evidence proving the controls were in place and effective. This was often a labor-intensive process involving screenshots, logs, and records which needed to be meticulously organized and maintained.
- Auditor Interaction: Interactions with auditors were typically more hands-on and continuous, as companies had to demonstrate their compliance through meetings, emails, and document submissions without centralized platforms that streamline these interactions today.
- Remediation: Any gaps found during the audit process had to be addressed manually. This often meant revisiting and revising processes, which could be time-consuming.
- Continuous Compliance: Maintaining compliance was an ongoing effort without automated monitoring tools. Companies had to regularly check their controls manually to ensure they remained compliant over time.



SOC2 has technical requirements that need to be constantly monitored such as MFA authentication. What are some examples of things that need to be constantly monitored in the context of ISCC? What are the things that need to be shown / managed?

- MFA not on all resources, Vanta is integrated with all our systems and identifies when that isn't the case.

Move your GRC program from point-in-time checks to continuous and automated.



Going /against <https://www.ibm.com/products/openpages>

Yc application:

I dont like company description. "Made easy" what does it mean? What does the company do?

Why turbotax ? why not sth like Vanta

Rimba is a software company to help manage sustainability certification compliance for manufacturers, starting with the renewable fuel market

Carbon carbon accounting software has gained crazy momentum in the last few years. Notably, both Watershed and Greenly have announced growth rounds in the last 60 days:

- Watershed raised a \$100M Series C led by Greenoaks, Kleiner Perkins, and Sequoia. Company started in 2019, and Taylor Francis is cofounder/CEO.
- Greenly raised a \$52M Series B led by Fidelity. Company started in 2019, and Alexis Normand is cofounder/CEO.

Greenly:

The screenshot shows the Greenly platform interface. At the top, a banner reads "EVERYTHING YOU NEED TO MAKE AN IMPACT" and "Once your emissions have been calculated, take action". Below this, there's a section titled "Personalized action plans" with three buttons: "Engage your suppliers", "Contribute to certified climate projects", and "Join ESG Connect". To the right, there's a grid of three cards labeled "Hello Ubisoft", "Hello Arena", and "Hello Foodies", each showing a pending status and a progress bar at 0%.

Gadi: Plenty of major corporations have announced net-zero emissions targets. And while that alone won't be enough to move the needle on climate change, these targets have gotten the

ball rolling. Those net-zero goals are trickling down, too, influencing companies throughout the supply chain."

Naturally, all of this trickles down into the manufacturers in markets like Asia and LatAm too, bcs if corporates care about CSRD, assuming that generates even more pressure for suppliers to care about other certifications like ISCC

Tim: Correct, companies in EU and US are mostly concerned about carbon emissions (doesn't matter if they are manufacturing, bank or a software company). They are concerned about the mandatory disclosure they have to make about carbon emissions. That's why companies like Greenly, sustain.life, watershed have taken-off.

A niche we noticed is that aside from carbon emissions, regulators are also starting to ask disclosure of where the raw material is coming from, social and governance aspect of these suppliers, and carbon emissions from the transportation that is involved in the supply chain until the consumer. These are captured in certifications like ISCC. This has not really taken off yet, so we still have a head start.

Gadi: I saw greenly helps corporates push certification measures to their suppliers. If this works, they could slowly onboard the complete supply chain into their platform to help trace carbon footprint. I'd be curious if they found some success with this

The question Id ask myself too is: Is it the logical next step for companies like Watershed to start adding other certificates to their platform after they win CSRD? If yes, which certificates are they likely to go after? If the market opportunity for certificates like ISCC is obviously large, what are the reasons for why they'd choose not to get into that market? Are there some key differences between these markets that would make it such that if Watershed expands, they'd start losing focus because this is a different problem/market?

Tim: Both watershed and greenly are more focused as a carbon accounting platform

How does CSRD compare to other ESG disclosures?

			TCFD	NFRD	ISSB	GRI
Must report	General	Governance, strategy, risks				
Report if material	Environmental	Climate change	Strong overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
		Pollution	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
		Water	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
		Biodiversity	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
		Resources & circular economy	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
	Social	Own workforce	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
		Workers in value chain metrics	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
		Consumers and end-users metrics	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
		Affected communities metrics	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Weak overlap with CSRD
	Governance	Business conduct	No overlap with CSRD	Weak overlap with CSRD	Future overlap with CSRD	Strong overlap with CSRD

■ Strong overlap with CSRD ■ Future overlap with CSRD
■ Weak overlap with CSRD ■ No overlap with CSRD

Gadi: Greenly says they automate and streamline the data collection process, but they don't show how their product works. Do you know how they do it?

Expertise: I am still struggling with this. The founders have clear domain expertise, and understand the process of ISCC extremely well. However, they fell short in painting the complete picture for his market. Perhaps initially due to my bad questions, but I later discovered Watershed/Greenly as two companies that are going after the carbon accounting software and a few certifications related to this by the EU. I had previously asked Tim for more successful competitors in the space, as I was alarmed his only references were dormant startups, so him not sharing this initially worried me.

- Does he think Watershed/Greenly are not direct competitors of Rimba? Was he trying to not mention these because they are already too large?
- Is it the logical next step for companies like Watershed to start adding other certificates to their platform after they win CSRD? If yes, which certificates are they likely to go after? If the market opportunity for certificates like ISCC is obviously large, what are the reasons for why they'd choose not to get into that market? Are there some key differences between the market Rimba is targeting versus Watershed's such that if they expand to it, they'd start losing focus or be at a major disadvantage against Rimba?

CONVOS WITH CHATGPT

The data requirements for ISCC (International Sustainability & Carbon Certification) tend to be more granular and specific to the sustainability of biomass and bioenergy products, focusing deeply on the environmental impacts and traceability of raw materials. In contrast, the CSRD (Corporate Sustainability Reporting Directive) covers a broader range of sustainability data, encompassing environmental, social, and governance (ESG) factors applicable across all operations of a company. Here are examples of the types of data required for each:

ISCC

1. Source and Origin of Biomass:

- Type of Data: Specific details about the biomass source, including geographical location, type of biomass (e.g., agricultural residue, specific crops, waste materials), and the method of cultivation or collection.
- Purpose: To ensure the biomass is produced or sourced sustainably, without contributing to deforestation or biodiversity loss, and to verify compliance with legal and sustainability criteria.

2. Greenhouse Gas (GHG) Emissions:

- Type of Data: Detailed calculations of GHG emissions throughout the lifecycle of the bio-based product, from cultivation and harvesting of biomass to processing and transportation.
- Purpose: To document reductions in GHG emissions compared to fossil fuel benchmarks, supporting claims of environmental benefit and compliance with sustainability goals.

3. Sustainability Practices:

- Type of Data: Information on agricultural practices (e.g., use of pesticides, water management, soil conservation practices), labor conditions, and compliance with social standards.
- Purpose: To ensure that the production of biomass adheres to environmental and social standards, promoting sustainable development and safeguarding worker rights.

CSRD

1. Environmental Impact:

- Type of Data: Broad data on energy use, water consumption, waste generation, recycling rates, and direct and indirect greenhouse gas emissions across the company.
- Purpose: To assess the company's environmental footprint and identify areas for improvement in resource efficiency and environmental stewardship.

2. Social and Employee Matters:

- Type of Data: Information on employee diversity (e.g., gender, age, ethnicity), labor practices, health and safety statistics, and human rights considerations within the

company and its supply chain.

- Purpose: To promote social equity and worker welfare, ensuring that the company's practices align with societal values and labor laws.

3. Governance Practices:

- Type of Data: Details on the company's governance structures and practices, such as board composition, executive compensation, anti-corruption and bribery policies, and shareholder rights.
- Purpose: To provide transparency about how the company is managed and governed, ensuring accountability and alignment with the interests of stakeholders.

The ISCC's requirements are more focused on the specific environmental and social impacts of biomass production and use, requiring detailed tracking and reporting to ensure sustainability throughout the product lifecycle. In contrast, the CSRD requires a wider range of data, reflecting the company's overall impact on sustainability, including governance aspects. Each type of reporting serves different purposes and targets different aspects of sustainability, with ISCC being more narrowly focused and CSRD encompassing a broader corporate perspective.

Understand workflow

What the product does

What does their product do?

Manufacturers have to go through thousands of docs (PDFs) every week. Docs are either (1) certificates of weight (weights of raw materials they receive), (2) contracts with their suppliers, and (3) delivery notes.

These manufacturers already track all of these for their manufacturing reports. The docs are in paper format, so they can scan them, and info is used for general ledger. If companies are not certified, they only care about the weights of each shipment, but if they are certified, they need to start tracking a lot more information such as:

- Traceability of raw materials
 - Where did they come from? Factory's location? Overlap with deforestation areas? Labor practices?
- Track real gas emissions
 - Calculate distance, type of trucks, how long of a journey, how much weight on trucks do they carry, depending on type of truck, type of fuel they use

Docs today are manually read, scanned, and added to multiple spreadsheets that track the aforementioned stuff. With all this data, companies create a mass balance spreadsheet that tracks all their inputs, and whether they are certified or not. This dictates how much % of their sales can go into markets that require certifications.

For example, if only 30% of their input materials are certified (in blended values) because they only manage to get 30% of their supply chain as certified. Then, only 30% of their sales can go

into those markets that require certification (e.g., Europe and Japan). They can cross this limit, but it's a loan to future certified materials, so in aggregate, they can't cross this limit. If they do, their certification gets suspended.

The mass balance helps track this, how much certified material that comes in can't be more than the certified biofuel coming out.

Things Rimba is building to help streamline this:

1. Data Extraction and Reporting Automation:

- Process unstructured data e.g. scanned PDF, purchase order, weight certificate, shipping documents.
- Automates data extraction for compliance reporting; Intelligent Document Processing (IDP)

2. Material Tracking and Risk Management:

- Trace material flows and segregate certified from non-certified materials.
- Prevents double accounting and maintains integrity of supply chains.

3. Streamline Multiple Sustainability Certification Audit:

- more transparency + less manual data entry + less manual calculation on excel
- Automates cross-certification, reducing effort and duplication.

Further breakdown of requirements:

What makes them special?

How do they convince people to work this way?

Rimba is focused on X, while Watershed is focused on Y. But it does help demonstrate that there is a market here because of Z.

▼ Case Against Bloated MVPs

I used to constantly fall prey to this — especially when working on projects in college — where talking to consumers in depth or deep rooted industry knowledge identifying a real gap was present. So I'd build products/tools that I thought were "cooooo" and "useful".

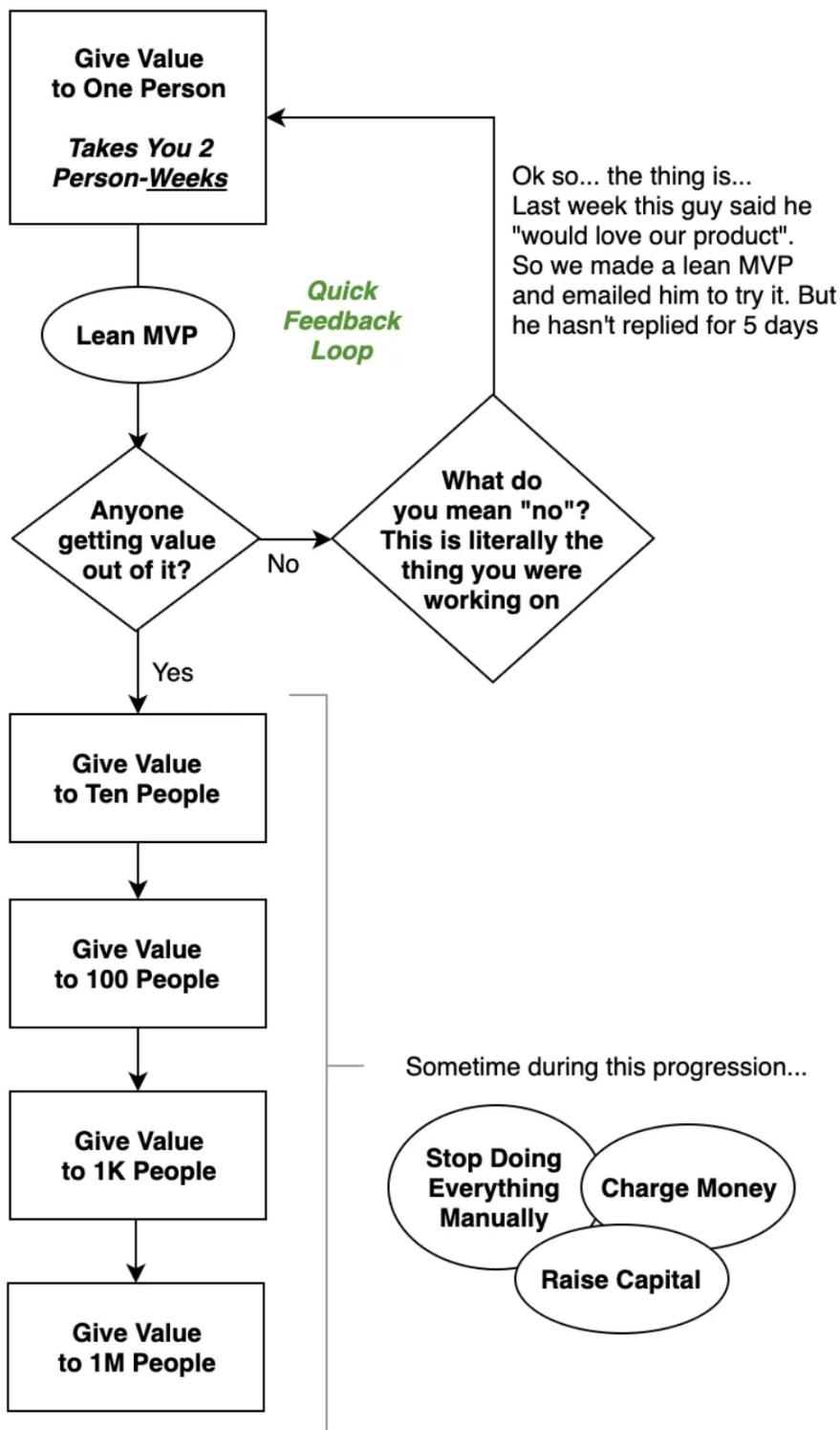
Send it out after so many hours of iteration. Product is done. You're on a high... the random people you sent it out to stop coming to the sites. There are no users and the website is a ghost town.

It was a bloated MVP!

So below are my lessons:

- rephrase the word Minimum Viable Product with Minimum Viable "Thing"
 - what's the thing that solves the solution - value

- i believe the commoditization of engineering or product building has left the ability to build insanely "cool" looking tools, without the necessity of the product existing — thus "bloated mvp"
- "is anyone getting anyyy value out of it? or do you have 0 users?"
- think of your #1 most compelling value prop story for your MVP
 - tell the story with as much **specific context** as possible
- generate specific value to 1–10 specific people you know, consistent with your specific value prop story earlier
 - the logo/ux can come in later
 - the aim of an mvp should be how can i get the first user and give them enough value to convince myself of the product i want to build



- coding the saying into an algorithm use a greedy search algorithm of giving value to someone. where maximum value can be generated is where entry should be done for even slight doubling down of features → further sales → a product → money etc...

- what's the most manual way i can give value to a person, then when it's actual value, i can scale it to a product
 - be a manual value consultant in your area — until you can't scale it anymore because of need → then go on building the mvp

if all this is true, why are we then still building bloated mvps?

- **Value Prop Blindness:** You don't really understand what problem you're solving, and you don't realize how important it is to pass this sanity check before building anything
- **Cargo Culting:** You want to build up your self-image as a "founder", and you have a mental image of founders building products, so you set out to build a product
- **Social Permissivity:** The startup community hasn't yet picked up on the idea that a Minimum Viable Product typically shouldn't be an actual product, so you get to plow ahead in the wrong direction without feeling socially pressured by your startup-peers to course correct until it's too late
- **Sense of Control:** Working on product design and engineering makes you feel (wrongly) like you know what you're doing and you're making tangible progress
- **Fun:** Product design is fun. Engineering is also fun.

▼ Improving Series A Working Capital

Background

Founded by the previous founders of Wunderlist (Acq. MSFT \$35m), Pitch is a productivity tool for optimizing presentations. Pitch has raised \$52m in equity funding series B that's primarily building for early adoption and product-market fit. They are planning to enhance global presence and grow internationally (aside from the US and Germany). I have interacted with the company when exploring the Berlin startup ecosystem and Contrary Capital's network(summer 2020 internship).

For context, Microsoft acquired Wunderlist without the will of the early founders who didn't have enough votes on the board. The serial entrepreneur Reber wants to minimize the unwanted costs and changes in ownership structure external funding will present by pacing/minimizing venture rounds. Given they have already taken out some debt from German banks, Reber wants to reduce debt financing from eroding operating margins.

Ok, let's think...

Pitch's customers pay subscriptions for individuals within their teams. However, to negotiate better deals that improve early adopter satisfaction, payment terms are done on a case-by-

case basis. To widen the period cash stays at hand, we have to make sure Pitch's \$10/member subscription fee arrives promptly. Pitch can improve this AR day by:

Setting up a direct debit process whereby recurring customers' payments are automatically collected on consistent days of every month. A direct debit mandate gives authorization for banks to transfer customer payments that streamline the collection process on both ends.

Most of Pitch's customers, being other startups working on their presentations, are prone to be late on their payments. To smoothen the AR gap, invoice factoring that finances open invoices from "slow-paying, but creditworthy customers" could be utilized. Invoice factoring would accelerate the deposit of funds to Pitch's account through an advance and an eventual rebate with rates as low as 1.15%. More importantly, the process is available to small businesses without credit.

Due to the relative discount price, users might sometimes opt for paying annual subscriptions ahead of time. Although this deferred revenue serves as a liability on financial statements, it's cash at hand - improving the WC.

Finally, incentivizing faster payment by offering discounts from 2% of the sale or an extra free Pitch.io account for early settlement will reduce AR days.

Although it's harder for early-stage startups to get longer AP days, the founder's past company Wunderlist's track records show they're "good for the money." Thus longer pay periods could be negotiated for cloud hosting AWS, legal fees, and salesforce expenses.

Like most software startups, Pitch doesn't need to consider the inventory lever when optimizing for working capital. A high-quality centralized source that supports customers - not products in stock - is required for supporting customers.

Working Capital Ratio	Scenario	Implication	To-do
Below 1 - Deficit	After optimization WCR = 0	There's a bigger issue (not only lack of optimization) Inefficient sales funnels Risk bankruptcy	Improve marketing and sales conversions by either shifting strategies or investing more money. Consider financing options (get debt) Debt restructuring
1- Breakpoint	Stable state	The business is a float but not enough for permanent WC Risks in emergency preparedness, sudden growth and expansion opportunities	Permanent WC is needed to pay debt and minimally continue core activities without interruption, so further optimization is suggested.

Above 1 - Sustainable	Enough WC to exist and thrive, but not overflowing to growth	Operational efficiency Higher company valuation Attraction of potential future investors (next round series C)	Good for mature startups that have stabilized churn and resurrected users. Pitch needs to grow and achieve product/market fit, thus sustainable stagnancy should be avoided but always maintained.
Above 1 - Surplus	There's enough permanent WC to operate the business, plus overflow.	Opportunity cost if there isn't an immediate function for the cash at hand. If growth is expected, it could be funded within the company - the surplus would fund the growth (growth capital)	If the WC is sitting there, reinvest or put the money in an active environment that beats the time value of the assets. (consider growth opportunities) If growth is expected/wanted, further fine tune the surplus to enhance internally funded sustainable growth. More money flowing from the WC towards the Growth Capital fund

It's common for some established tech companies to have their current ratio under 1, liquidating brand assets to usable WC. Although Pitch's founders are known, the startup is still young and shouldn't risk a negative working cash flow. Especially with the current pandemic, as startups have gone "under" due to an overestimation of brand value and liquidation failures.

Suppose the Pitch has done all the optimization strategies mentioned above and is still running on a ratio below one or hasn't produced surplus WC for Growth Capital. In that case, alternative financing options have to be considered.

Given Pitch's track of paying off debt, innovative startup credit cards like Brex could be used if there's an expectation of fluctuating profits or insurance for unexpected WC drops. Brex cards can smoothen the unexpected drops as there's an inconsistency in AR days to onboard high-value clients that enhance traction but demand more extended periods.

Alternatively, Pitch could access innovative early-stage startup financing options, including cash flow loans and invoice financing at Fundbox and Blueville, that allow borrowing against outstanding invoices due.

Given Pitch has taken out multiple loans from different banks; debt modifications should be employed. Visiting their BayernLB bank Pitch should refinance for lower payments or consolidate various loans to replace them with a single, more affordable loan.

Growth

Given Pitch's unwillingness to give more equity till the next round of fundraising (series C), the ballyhoo of "growth at all costs" doesn't work. Previously raised equity capital's elongated runway strategy remains intact as the company explores alternatives used for additional growth opportunities. Growth should be with internal WC (the part that's overporing to the Growth Capital fund) at an SGR.

Advancing, Pitch should use the cash-flow sustainable growth rate instead of pure sustainable growth rate to take into account cash flow generated by the operational activities. Given our limitations of WC, we can further enhance and fine-tune the surplus WC (option 4 on the table above) by continuing to:

- (1) shorten the operating cash cycles with the AP/AR modifications,
- (2) increase the profitability ratio by fine-tuning every expense
- (3) reduce the asset turnover ratio by increasing the threshold gap needed before hiring additional people and resources. Plus, "testing-the-waters" before taking the financial risk of global expansion is necessary.

Keep a consistent, sustainable WC ratio that adjusts for unplanned needs. After determining the target WC ratio, we have the freedom to select a growth rate that corresponds to the amount surplus from the permanent WC requirements (shown in: Above 1 sustainable in table 1) that pours onto the growth capital bucket. To maximize the results, use the entirety of the surplus (Above 1 surplus (minus) Above 1 sustainable = Growth Capital GC) and set that as Pitch's SGR. Going further, Pitch can use this [working capital calculator](#) to determine the WC requirements every month and stay on track of achieving the US market expansion growth objectives.

▼ Approaching Ideal Money - Are Non-Collateralized Algorithmic Stablecoins Feasible?

<https://prod-files-secure.s3.us-west-2.amazonaws.com/c4c0fb6a-3b8e-44a5-b1c4-2c5efdb2d812/6fa12ce1-0640-44da-b268-5343122f334c/Paper.pdf>

Arts

▼ Phoebe Philo

Phoebe A"cheaper"1 Philo

Quoting @yosymichael from Twitter: 'Phoebe Philo didn't stop at "Open your purse!" She said, "Sell your house, rob a bank, and do some credit card fraud too!"' When the long awaited email dropped, I'm sure some tumbler age "Phiophites" gasped.

phoebephilo.com is now open

Inbox

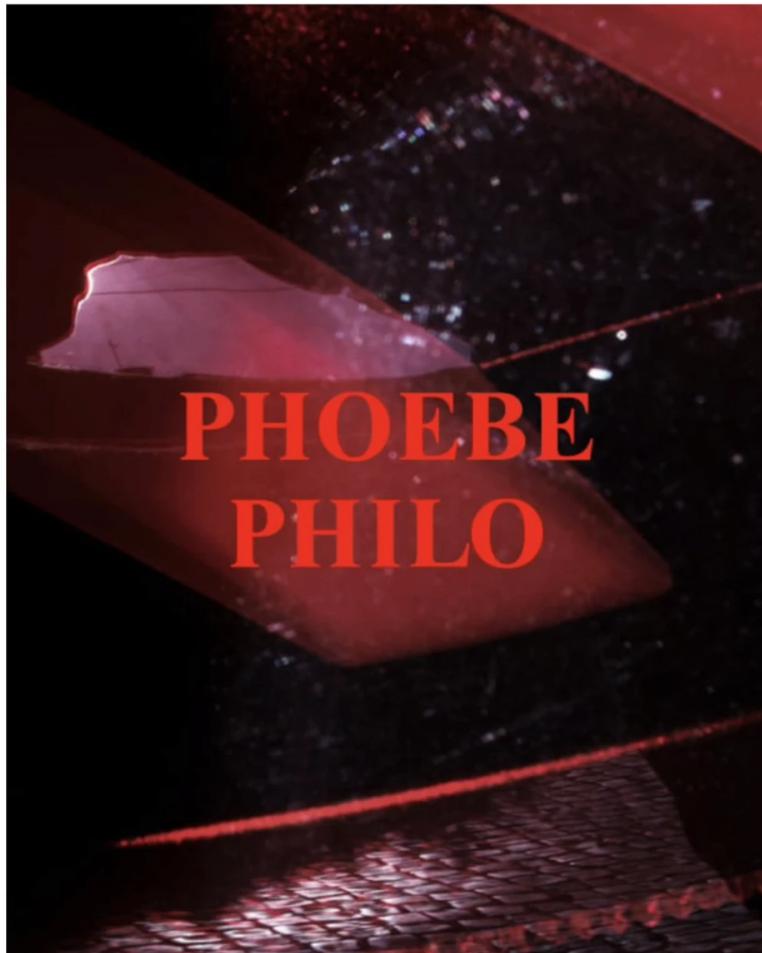


Phoebe Philo 15:04

to me ▾



...



[phoebephilo.com is now open](http://phoebephilo.com)

Background

To contextualize Phoebe Philo, we need to step back to "old Celine" and the legacy Philo had on "chic minimalism". she's the mommy of what's now TikTok cringe of "quite luxury".

Phoebe Philo, after serving as the creative director at Chloé for five years, left in 2006, succeeding her friend and predecessor, Stella McCartney, who had departed in 2001. Philo

carried forward the growth momentum initiated by McCartney, garnering a dedicated customer base. Her exit from Chloé, cited as a choice to prioritize her family, was unexpected given her career peak. (Glance at her namesake brand's MUM necklace)



Her absence led to the rise of brands with a similar aesthetic, like Victoria Beckham in Spring/Summer '09, The Row in Autumn/Winter '07, and H&M's COS in Spring/Summer '08. Though these brands emerged during Philo's hiatus, their establishment would've begun in her absence, with COS, backed by H&M, being swift due to more resources.

Philo's influence is undeniable; even lesser-known to the general public, her design essence impacted the fashion realm. She made a return, not with her own label but as the creative director at Celine in 2008. Her debut show was Spring/Summer '10. Despite challenges and competition, Philo's distinct, consumer-focused designs set her apart. Her time at Celine further cemented her influence, elevating the brand's financial standing in the industry.

Let's quickly go over what made "old celine" the golden days, and birthed a cult following.

I. The Hallmarks of Philo's Design Aesthetic:

Palette Choices

- **Monochromatic Mastery:** Philo often favored a muted, monochromatic palette – a deliberate choice that exudes a sense of timeless elegance.
- **Emphasis on Neutrals:** The use of beige, white, black, and navy became almost synonymous with her tenure at Céline.

Bookmark: Materiality & Texture

- **Tactile Luxury:** From the buttery leathers to crisp cottons, the materials scream luxury but in a whispered, understated tone.
- **Material Interplay:** Often paired contrasting materials, like wool with silk, to create depth and intrigue.

II. Silhouette & Structure:

Bookmark: Oversized Elegance

- **Effortless Oversizing:** Philo championed the oversized silhouette, proving that volume can, paradoxically, highlight femininity.
- **Tailored Fluidity:** Despite the ample fabrics, there was always a tailored element, whether in a cinched waist or a carefully draped fold.

Bookmark: Functional Femininity

- **Pockets and Comfort:** Her designs often incorporated large pockets, an ode to practicality without compromising on elegance.
- **Ease of Movement:** Flowing trousers, loose blouses, and drop-shoulder coats allowed for unrestricted movement.

III. Iconic Pieces & Collections:

Bookmark: The 'Old Céline' Trope

- **The Trapeze Bag:** A beautifully structured bag with wings, it quickly became an 'It' item under her direction.
- **Glove Shoes:** The V-cut shoe design, both in flats and heels, became a footwear phenomenon, emphasizing comfort and chicness.



Bookmark: The 2015 Spring Collection

- **The Modernist Touch:** Philo's play on proportions, asymmetry, and tunics over trousers presented a fresh take on layering.

- **Subtle Femininity:** Pieces like the knit dress with flowing strands heralded a new, confident femininity.



Her Protégés show how far back her roots extend. She nows she's Phoebe Philo and she birthed all of them

Daniel Lee - After working under Philo at Céline as the Director of Ready-to-Wear Design, he took the helm at Bottega Veneta in 2018. Under his leadership, the brand saw a significant rejuvenation in its aesthetic and became a favorite among fashion enthusiasts and celebrities.

Naza Yousefi - She was a former accessories designer at Céline during Philo's tenure and later founded the handbag label Yuzefi, which has become quite popular.

Peter Do is a notable designer who once mentioned that he was influenced by Philo. After studying at FIT in New York, he worked at Derek Lam and then joined Céline under Phoebe Philo, although he never worked directly with her. He later founded his own eponymous label, which is known for its tailored pieces and minimalist aesthetic reminiscent of Philo's work. He's now heading the comeback of Helmut Lang.

Rok Hwang - The founder of the brand Rokh worked at Céline under Philo. His label showcases deconstructed pieces, precision tailoring, and unique details that hint at his experience under Philo's mentorship.

Lucie and Luke Meier - While their direct connection to Philo is not as former direct subordinates, they've exhibited aesthetic affinities with her. The duo, currently at the helm of Jil Sander, bring a minimalist and thoughtful design approach to their collections.

Gabriela Hearst - Although she did not work directly under Philo, Gabriela Hearst's design ethos, which is sustainability-driven with a minimalist touch, has often been compared to Philo's work.

New Collection

The elephant in the room - Price

Let's use the brand's most expensive bag currently, the 'XL Cabas, as an example for how baffling Philo's pricing is. It's a huge tote bag in calf leather—"calf leather," remember that--that's selling for \$8,500. Now let's go back to the mid-2000's where an average bag from a fashion house can costs about \$700-\$1,500. Philo's famous 'Paddington' bag in leather when she was at the creative head of Chloe was around \$1,500 at the time. Even the most expensive bag of a fashion house would always be made out of exotic leather, that would've costs around \$4,200.

The price was a positioning necessity. Through it, Philo placed herself at the upper echelon of fashion — both in brand positioning and price. It feels like LVMH's attempt to have an uber-luxury (Channel and Hermes. To be fair, it's the price-point with the higher growth potential and less saturation. Hermes trying to keep the LVMH size just at 17%, so they're trying to go around. And although Phoebe Philo doesn't have the same demand power due to lack of "heritage" — and it's hard to find a designer that carries the same amount of weight Philo.



"Affordable Luxury" is sort of a weird thing to say for people that know Phoebe. These commenter are unserious and should really stop weighing in on these subjects. As @shannon_sense put it, "even if everything was reduced by 30% it would still be too expensive for most people." There's sooo much to be said about the ways people evaluate luxury fashion from their own perspectives rather than from the perspective of luxury customers.

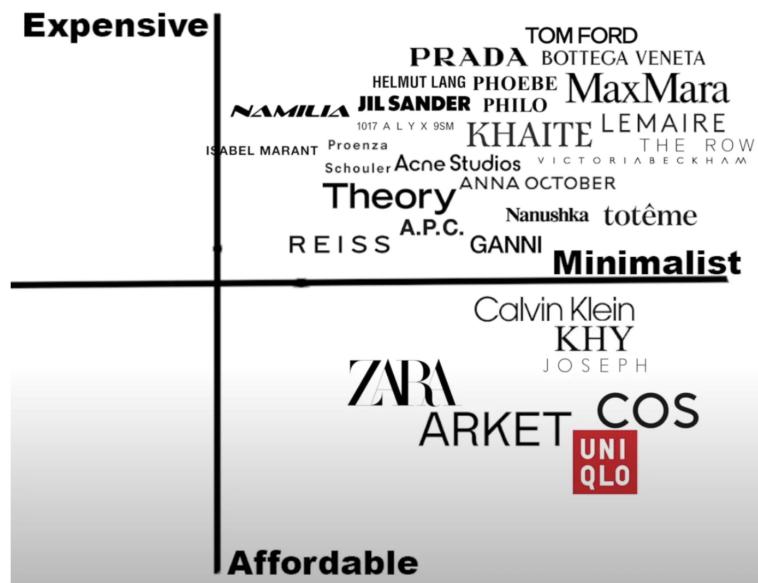
I disagree with people that say it's extremely expensive. It's not. At least for Philo's perceived standing. Chloe and Celine customers might not have prepared for the astronomical prices. I was expecting a higher price. They're more mixed than I expected. I really was anticipating only the high priced items at first to establish the price point of the brand. But the range was a pleasant surprise, entry point products are usually a good idea and allows most people to get in on the action. I just didn't expect so many so early.

Look, i'm not saying I can afford this, but I'm meant not to. It's meant to be something to look forward to and them achieve. Not just buy. Pricey, but its a well studied price point, so I think its smart for the branding and positioning

However I disagree with people that say, similarly to Daniel Lee at Burberry right now, Philo must first acclimate her old customers used to her Chloe and Céline prices, as well as her new and potential ones to this new pricing under her name so that the bridge between "want" and "closet" which is actually buying the pieces and ensembles becomes easier to walk on.

Also between the time that I started writing this, and now finished, over 50% of the products in Philo's website have sold out. Her target audience are ready to spend their money for her and that's what she focused on. Touché

That leads me to this Rabkin post reacting to the initial release announcement about the mythical nature of Philo and luxury at large.



Phoebe Doesn't Exist

This aura of this collection just selling out without any ads, runway, posters, influences, or anything the typical "high-fashion" world is used to, reminds me of Eugene Rabkin's blog "Philo doesn't exist". Quoting directly, "the collection will be revealed [and it was] not in real life but through images – simulacra – and will be sold online, a hyperreal way of shopping. No one will

have any direct contact with the clothes – arguably the only piece of reality here – until they will get a box at their home. Until then, no one will know how the materials feel, how the garments fit, or their true colors. We will not get an insight into Phoebe Philo's work process, because she does not give interviews. We will never really know who designed the collection, how it was designed, and what it really looks like. The entire thing is a simulation.



To drive his point further, in 1991 Baudrillard wrote three articles about the first Gulf War: *The Gulf War Will Not Take Place*, *The Gulf War Is Not Really Taking Place*, *The Gulf War Did Not Take Place*. Of course he did not mean that there was no military action happening in Kuwait; what he argued was that our only experience of the war was through a narrow channel of highly mediated messages that have only tenuous relationship to the reality on the ground.

In other words, we live in a simulation – via screens, through social media, soaked in a semiotic system created by the vast leisure industry – entertainment, news, advertising, and so on. Similarly, when the Phoebe Philo collection came out this fall The Gulf War did not take place. Phoebe Philo does not exist.



the shoes are hot!



Right Time?

In a 2006 statement responding to creating her own brand, she said it wasn't the right time was actually right on the money so now that we are once again in an awful financial situation with what seems like an impending recession, the good old minimalism rises, and to Philo, this seems like the "right time" — along with Arnault's backing ofc. Looking at the correlation of fashion and recession,

- Chanel having a rise around the Depression in 1920s.

- 1991 Helmut Lang and Jil Sander
- 2008 recession and Philo's rise at Celine. At the time, she was sweeping away the excess of aughts fashion with a confident new minimalism. Tapping on similar instincts now, she has an even bigger following to rely on.
- In 2023, the cringe "old-money" and elevated basic.

The messaging from the website seems spot on: "Our aim is to create a product that reflects permanence."

Subtle Wardrobe Direction

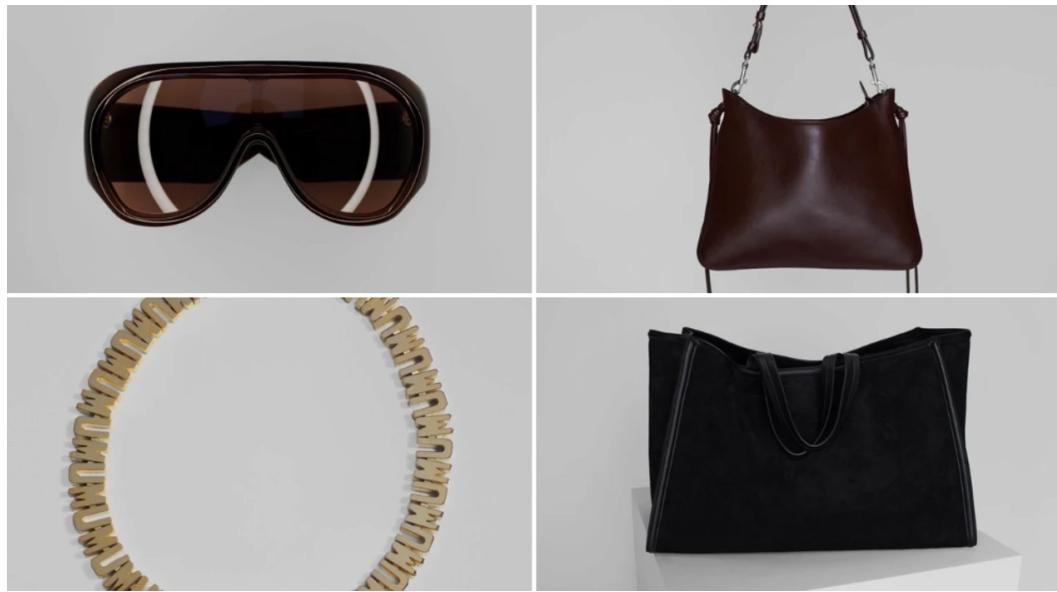
This Phoebe Philo feels like The Row had an affair with Rick Owens and the gay son was Bottega and that daughter was Loewe very chic.

Everything on the website seems relaxed, less trendy, dignified. It's reminiscent of older couture when designs were for women over 40, and aspirational. I love that she's separating the girl from the women. This collection her news collection is about women. No Influencers, no celebrities just design and great products! The most important thing about it is intrinsic value. So much of fashion is what other people like And not what the customer actually likes. Now. It's swinging back to the customer, and I'm all for it.

The ultimate modern wardrobe from **a dissatisfied woman**" says it all. And I like the collection...with caveats. But what really excites me is the idea that this collection could...maybe...free other designers from the crushing cycle of, as Horyn puts it, "chasing growth." That chase has literally killed some of our greatest modern designers, and driven others to breakdowns. If this new Phoebe Philo augurs a new model, I'm all for it.



Rather than attempt radical change, Phoebe Philo's new collection offers women a subtle way to evolve their wardrobes. Having pushed boundaries before, Philo understands most women reach a point where overhaul is replaced by nuance. Adding special pieces allows self-expression, not reinvention. Philo knows women harbor hidden boldness behind practicality. Her clothes enable this duality. Witness trousers with a teasing back zipper, or a toothpick pendant necklace for discreet utility. Philo relates to the life stage where less becomes more. Her "edits" resonate by providing the special over the sweeping. Limited availability complements crafting a wardrobe across seasons, not discarding it each time. For women seeking expression through subtlety, Philo provides the perfect avenue in this new collection. Its allure is in Evolutionary, not revolutionary, dressing.



"And my chick in that new Phoebe Philo

So much head, I woke up to Sleepy Hollow" Ye

▼ Paris Spring '24 Men's Quick Reflections

The most rick a rick show has been since pre pandemic.



Uncertain future at Lanvin is the re'see the new show?

Prada slime. prada, shorts?



Wait 032c makes clothes? Yes, and you're late.

Louis Vuitton — what's a king to a god, what's a show to a spectacle, what's a spectacle to JAY-Z.

- I liked the role it had during the LV show, but also questioned the deviation between playing with actual great design while embracing hip hop (i.e, Fear of God) vs this (which seemed slightly off tbh).

lol pusha t's coke music being played for so many of the men's runways

It's the year of Jonathan Anderson.

Gucci has to be changing entire comms strategy by Sept.

Zegna as the less cool Fear of God!

The Row moved all operations to Paris

It's Rhude to owe people money lol

Los Angeles Times

CALIFORNIA

Rhude streetwear founder raided company to fund lifestyle, lawsuit says

Lemaire is great i wish they talked more!

Did anyone talk about Saint Laurent

Why would you wear Hermes RTW as a man when there is Loro Piana

A magazine curated by sacai - the bluest blue

Everyone is on ozempic fr - empty box in their fridge at posh hotel Le Bristol



VIEWPOINT

I definitely would have taken Ozempic to lose weight five years ago—but what else would I have lost?

What would we do—who would we be—if we let our bodies and lives exist more or less as they are, without looking to drugs like Ozempic to free us from the eternal trap of bodily perfection? Emma Specter unpacks the alarming craze.

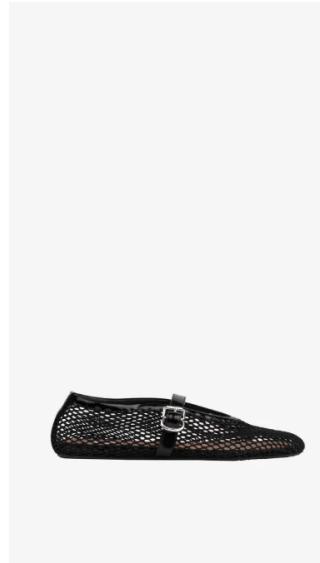
BY EMMA SPECTER

Does anybody read boring reviews?

If you walked for Junya i hope your hair is doing ok!



The aldia net flats are the shoes of the summer



Kenzo aw man...

Jacquemus dropped the ball

▼ Glenn Martens - Vanguard of the Modern Silhouette?

Glenn Martens: Vanguard of the Modern Silhouette?



Three nuanced ensembles from Glenn Martens, shot by Luis Alberto Rodrigues. Showcasing his work across Diesel, Jean Paul Gaultier, and Y/Project.

Martens' rapid rise illustrates how hungry the industry is to anoint a new savior. He's trying to bring back avant-garde experimentation to the high-fashion mainstream, after years of minimalism dominating luxury fashion. Yes, he modernizes Y/Project and Courrèges with a youthful energy. What I worry about is: does his seemingly recycled aesthetic lacking in true creativity? Wondering if he's overhyped and underwhelming given the industry is too eager to coronate the next big thing before they are ready.

Drawing inspiration from the fluidity of architecture, Martens creates designs reminiscent of the natural mountainous vistas of his homeland. It comes as no surprise that he shares the podium with JW Anderson, another proponent of sculptural fashion. Both trailblazers, Anderson leads LOEWE and JW Anderson, while Martens' indelible mark is felt across multiple maisons.

Manifesto of Martens' Aesthetics:

- Layered intricacies.
- Thoughtful twists.
- Bold prints.
- Precise folds and jagged edges.
- The art of the oversized and asymmetrical.

In a world obsessed with commodified luxury, Martens tactfully navigates. With a touch of pragmatism, he merges forward-thinking experiments from Y/Project with Diesel's vast outreach. The result? An exploration that satisfies both fashion elites and mass markets.

L'Artiste's Choice?

Y/Project, under Martens, strikes a chord even in its offbeat notes. While some ensembles soar, others provoke thought, but each carries Martens' unmistakable signature. An echo of the late Yohan Serfaty's vision, Martens has ushered in a new era. He juxtaposes Serfaty's legacy with his lavish, fun, and slightly audacious style. Influenced by occidental aesthetics, Martens has iteratively evolved the brand, deftly playing with proportions, silhouettes, and fabrics.

Y/Project pulsates with a unique cadence. Sometimes it's harmoniously in sync, while occasionally it stumbles. But there's undeniable novelty. As aptly described by "Fashionlover4", it's akin to a "fashion student's wet dream". While Yohan Serfaty laid its foundations, Martens has expanded its horizon with gender-fluid, enigmatic designs evocative of the iconic Rick Owens.

Resurgence of the Denim Giant

Diesel's renaissance under Martens is nothing short of remarkable. Martens dips into the brand's golden era of the '90s and early '00s, stirring nostalgia while redefining Diesel's modern identity. The brand's playful metamorphosis from "For Successful Living" to "For Sucsexful Living" post Martens' intervention is emblematic of his audacious touch.

Yet, as Martens flirts with Diesel's denim legacy, there are challenges. While the runway flaunts luxury, the racks sometimes reveal impractical flamboyance. The harmony between Diesel's essence and Martens' flair needs fine-tuning. But, given time, there's no doubt Martens will blend his experimental spirit with Diesel's rich denim history, offering fresh takes on classic staples. With a dash of patience and Martens at the helm, fashion enthusiasts worldwide can anticipate a reimagined denim dynasty. The future might be denim, and it's couture.

His exaggerated proportions and gender-bending styling cover familiar ground already charted by predecessors like Martin Margiela, Ann Demeulemeester and Raf Simons. But, unlike those designers' radical conceptual garments, Martens' oversized blazers and trench coat dresses are tame in comparison. Yes, his subversion of masculine and feminine codes challenges binaries. But in 2022, gender fluidity in fashion is practically mainstream. Younger talents like Harris Reed and Telfar Clemens are doing more to expand definitions of identity and expression through clothing.

Where Martens does excel is in his digital-print tailoring and knitwear. Martens once shared, "My design realm is vivacious, whimsical, and a tad provocative." He honors the brand's essence while continually exploring new forms and dimensions, mirroring an artist rediscovering age-old hues. Pieces like the anatomical prints from the spring 1996 "Cyberbaba" collection exemplify his masterful reinvention. His pixelated and blurred suiting fabrics, often in bright hues, have a hyper-modern vibrancy. The asymmetrical color-blocked knits he designs are actually quite imaginative in their use of graphic color and texture. Martens clearly has an affinity for digitally manipulated textiles and colors. When applied to Y/Project's signature oversized tailoring and body-hugging knits, the results bring an edgy, hypermodern look to life. His custom fabrics point to the potential he has to develop a more distinct design identity.

Looks

Spring 24

I like how he's redefining what wearability could look like. The issue I constantly face is when he doubles down on playing with the structure and color. It's overwhelming and far from pleasing.

- Highlights:
 - I've fallen in love with the buttoned boot





Later end of the show consisted of fabrics that look like they've been paper-mache style dried immediately after the washer. I'm a fan of the tops, skirts, and pants in this texture. Including denim!



Detail Look 118



Detail Look 119



Detail Look 120



Detail Look 121



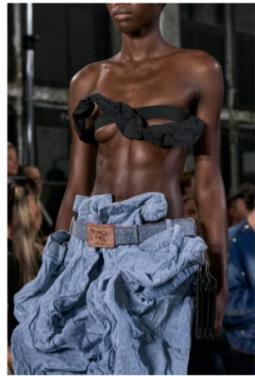
Detail Look 122



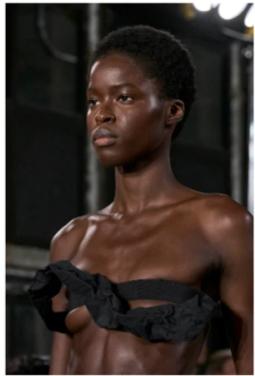
Detail Look 123



Detail Look 127



Detail Look 128



Detail Look 129



Detail Look 130



Detail Look 131



Detail Look 132

Fall 23

- This feels like the official incorporation of denim bleeding from Diesel into Y/Project's work.
- Highlights:
 - Denim bleeding into other fabrics in a tasteful manner



Look 31

Look 32

Feels like not knowing where the clothes end and the shoes start

▼ J Hus

J Hus

Go listen to this guy. A true generational talent that's at his peak right now. I can't think of anyone that's at this potential level at the moment.

I feel personally attached because of the following reasons:

- There was a time where my most played genre was UK Hip Hop
- I'm Ethiopian and love the traditional African sounds/instruments
- I love it when words are emphasized and you can not only hear the words but feel them

But he's had a three album **RUNN** now;

1. Common Sense

"Came in a black Benz, left in a white one I'm just a hoodlum I came with bonsam"

2. Big Conspiracy

"We run from 'rales with the mullianis They can't see my face, I'm like a hijabi I gotta stack bread 'cause I'm building my army They know I'm so solid, they callin' me Harvey I get all the 'usna and all of the narnis"

3. Beautiful and Brutal Yard (BABY)

"He weren't the same when I saw him again, he's a real shapeshifter Used to pray facing qibla I just chill in my sector, you know I'm Hannibal Lecter Put that boy in his Pampers, us man, we're

not rampers Post outside, we're campers, come to your uni campus Splash him, John the Baptist, I don't need no accomplice Maybe only a driver, turn that man to a diver That day, it was raining, put on the windscreen wiper"

Also unrelated to Hus, but general music commentary – it's truly magical how three biggest artists in the world can combine forces to make a song THAT boring. Hopefully Utopia isn't that bad.

▼ Invisible Hand Actually Malevolent?

If you're not familiar with the original [SSC essay](#), read my [summary below](#) before reading my thoughts on top here...

Invisible Hand Actually Malevolent?

A Review of Meditations on Moloch!



Moloch is about the triumph of incentives over values. The triumph of instrumental goals over terminal goals. The Nash-Equilibrium where the system is at a steady state is Moloch. The source of most evil. A trap where people can't get out of as they are forced to think and act locally. Falling prey to the competitive forces that maximize individual outcomes, instead of preferring cooperation to submit to the god of our values. Moloch appears at any point when multiple agents have similar levels of power and different goals. Moloch exemplifies unfortunate competitive dynamics.

Deep down, nobody actually wants it to keep going this way, even the winners. It's a hedonic cycle for civilization. Left unchecked, it will sacrifice all our values and all we really value. "Sacrifice values to get ahead." It is not necessarily greed; at points, "getting ahead" becomes necessary.

"Coordination problems create perverse incentives" is a very basic tenet of economics, which is essentially what the post boils down to. However, this economics-101 sentence is dull, uninspiring and doesn't really tell the entire story. Scott Alexander takes a perhaps poetic way

of introducing the concepts to those who are unfamiliar with them. Mr. Alexander is a lecturer who had jazzed up "Week 4 - Coordination Problems" with a poetic personification, but with little economics literature around such problems. To do so, Alexander uses Allen Ginsberg's poem, which serves as the post's underlying theme and is referenced throughout. Even with my familiarity with the concept of coordination problems, I still thought the poem itself was esoteric. I don't think referencing the poem helped to explain the concept. From the surface, it leaves the impression of writing things that sound intellectually rigorous as opposed to writing something that is actually intellectually rigorous. For the most part Alexander avoids this, but the Moloch stuff is more dubious.

In Ginberg's poems, Moloch isn't just a literal god. Neither a set of equations. Moloch is part of human nature — one we're horrified by. Scott Alexander does a good job of building the image of Moloch in our world. It gives off a vague, yet powerful sense of knowing. It sort of allows one to have a shorthand answer to why things happen — Moloch!" What is Moloch? The demon god of Carthage, and to him we say Carthego delenda est"

Where do we go from here? Per SSC, to defeat Moloch, we need an agent that we side with holding human values. "Elua" or the "Gardner" that will optimize for what we like. The essay reads ominous. Scott takes Ginsberg's poem and retells it — nature has fucked us over, and reason is the only thing that can save us from it. This reminds me of Bucky Fuller's quote "You never change things by fighting the existing reality. To change something, build a new model that makes the existing model obsolete."

Alexander's bias is along the lines of "AI is the looming existential threat that will kill us all". The first AI to hit Singularity-level will outstrip everything around it in terms of intelligence, and so would truly be a singular entity with no competition. This seems, to Alexander, not just a utopia, but the only viable way of escaping the Malthusian trap. I'm assuming this relates to good superintelligence – the only thing that will save us from a bad one, is a good one that sides with us. A battle between the evil god Moloch, and an alternative god Elua — a superintelligence that has values aligned with humans.

It's tempting, and intellectually satisfying, to look at a set of problems, extract a meta-problem and then propose a solution: by solving the meta-problem, you solve all of its instances, too. However, the effectiveness of the solutions is dependent on how well the abstractions fit the instances. Plus, how unintended consequences won't overshadow the benefits. The singular autocrat may stop us from races-to-the-bottom, but can implement policies we're not particularly happy about.

In Alexander's case, he just wants a mechanism to stop competition inevitably sliding into local optimization traps, not necessarily advocating for an ideal utopia. Surely our super-intelligent AI overlord would be tempted to stray outside those bounds and look for other ways to help humanity out. The AI is far smarter than we are and has the wellbeing of all of humanity in its purview. How long until it decides that it knows with certainty that it can better manage our happiness than we can?

So, what then?

I guess for Marx, capitalism was Moloch, and communism was a solution. While the god-like powers of a super-intelligent AI could potentially solve Communism's information problem, it can't know what is in people's hearts. It will provide a target for the power-hungry to attempt to co-opt, and in defending itself is likely to crush the freedom and flourishing that it was supposed to nurture. There's a fatal flaw which has been demonstrated time and again by attempted instantiations of Communism: there are people who will go to unimaginable lengths to secure power. They outcompete anyone that's mild mannered, and eventually the whole system collapses. Although it's hard to predict how this will take place under our new AI overlord, I can predict it will happen ad-nauseam. Maybe the AI will detect and prevent subversions, but similar to autocrats' attempts, it's hard to do without clamping down on freedom in general.

Similarly, one might argue there won't be coordination problems if everything is ruled by one royal dynasty / one political party / one recursively self-improving artificial intelligence. To begin with, royal dynasties and political parties are not singletons by any stretch of the imagination. Infighting is Moloch. Getting to an absolute power required sacrificing a lot to Moloch during the wars between competing dynasties/political systems. But even if we assume an immortal benevolent human dictator, a dictator only exercises power through keys to power. Plus, has to constantly fight off competition for his power. Stalin didn't start the Great Purge for shits and giggles, and The Derg didn't assassinate literate and opposing politicians in Ethiopia for nothing; it's a tried and true strategy used by rulers throughout history. Royal succession, infighting within parties, and interactions between individual modules of the AI, all sacrifices to Moloch. The hope with artificial superintelligence is that, due to the wide design space of possible AIs, we can perhaps pick one that is sub-agent stable and free of mesa-optimization, and also more powerful than all other agents in the universe combined by a huge margin. If no AI can satisfy these conditions, we are just as doomed. Even then, there's the fragility of the outcome – there's a huge risk of disutility if we happen to get an unfriendly artificial intelligence.

For Unabomber, the method to stop Moloch was the destruction of complex technological society and all complex coordination problems. I categorize this solution in the primitive bucket whereby one assumes all problems will be simple if we make our lifestyle simple. But that's not defeating Moloch, but completely and unconditionally surrendering to Moloch in its original form of natural selection. Goals are mismatched. Avoiding Moloch is an instrumental goal; the terminal goal is to promote human well-being. But in primitive societies people starve, get sick, most of their kids die, etc. Additionally, this doesn't work in the long term; even if you would reduce the entire planet into stone age, there would be a competition to see who gets out of the stone age first – which got us here in the first place.

A lot of the rationalist community is focused on AI, which makes sense in that light of the existential risk of unaligned AI. However, looking at projects focused on non-AI solutions to countering or defeating Moloch, I ran across Game B. Game B seems to be a discourse around creating social norms that defeat moloch. So far it seems to me like a group of people who are trying to improve the world by talking to each other about how important it is to improve the world. "What are all those AI safety people talking about? Can you please give me three

specific examples of how they propose safety mechanisms should work?" I haven't seen easy answers or a good link for them.

Do Moloch and Eula co-exist? Aren't they one? An enforcer god(Moloch) for the prize (Eula). Would we want Eula's values if we didn't strive for it? Anyways, let's finish off with this beautiful deception by Dostoevsky on the pessimism of utopia: *"Shower upon him every earthly blessing, drown him in a sea of happiness, so that nothing but bubbles of bliss can be seen on the surface; give him economic prosperity, such that he should have nothing else to do but sleep, eat cakes and busy himself with the continuation of his species, and even then out of sheer ingratitude, sheer spite, man would play you some nasty trick. He would even risk his cakes and would deliberately desire the most fatal rubbish, the most uneconomical absurdity, simply to introduce into all this positive good sense his fatal fantastic element. It is just his fantastic dreams, his vulgar folly that he will desire to retain, simply in order to prove to himself--as though that were so necessary - that men still are men and not the keys of a piano"-- Notes from Underground



Summary of Initial Passage

Introducing The Beast

In Part I, the essay situates the main issue/character at play Moloch by illustrating him through Allen Ginsberg's Poem and multipolar traps that exist within society. In response to C.S Lewis' question "What does it? Earth could be fair, and all men glad and wise. Instead we have prisons, smokestacks, asylums...Sphinx of cement...eats up their imagination? The poem responds "Moloch does it" This part characterizes the theme of the essay by introducing us to Moloch -- the humanized version of civilization that we can almost "see". Through Bostrom's example of a dictator-less dystopia, Alexander introduces a lack of strong coordination mechanisms. From a god's-eye-view, we can optimize systems (especially ones filled with hardships with simple agreements, however, no agent within the system is able to "effect the transition without great risk to themselves".

To further illustrate these coordination issues, Alexander uses 10 real-world examples of multipolar traps: The Prisoner's Dilemma, Fish-Farming Story (one sneaky farmer will find a way to not pay for treating the shared pond, and the entire system follows), The Malthusian Trap (rats on an island are happy and "play music" until resources start being depleted by overpopulation becoming hard to exist, let alone play music), The Two-Income Trap (having a second job becomes the norm, without increasing quality of life if everyone does it), Agriculture is a less enjoyable way of living, but we are overpopulated so we need it, Arms Race (esp. expensive nuclear standoffs leading to heavy overspending of budgets that could go to better use), Cancer (only certain human cells overpopulating killing the host itself), and The "race to the bottom" where politics are pushed toward being more competitive than optimal for development of the society it leads.

Also other categories of multipolar traps where competition is regulated by an exterior source, i.e. social stigmas. Education - current methods are bad, but there is social signaling at play that perpetuates the system forward. Science - funding research, peer-reviews, and statistical significance tests are flawed, but rigor reduces the incentives a scientist gets from the previous mentioned methods. Government Corruption. Congress - "From a god's-eye-view, every Congressperson ought to think only of the good of the nation. From within the system, you do what gets you elected."

Questioning Our Motives

In this part Scott questions why as evolved and cognizant humans we fall to these traps. Answer – incentives hard-coded. Expands on why it's hard to switch these incentives. Due to these competitions everyone's "relative status is about the same as before, but everyone's absolute status is worse than before." Incentives drive us collectively and they're built in analogy of terrain to determine the shape of the river. Although building canals by altering terrains is possible, it's hard nonetheless. Incentives are hard to change -- especially from the hard coded ones of humanity. It's because of these incentives that things like Vegas, that doesn't optimize civilization, but "exists because of a quick in dopaminergic reward circuit", exist.

Retardants Of Our Downfall

Given the beast and our inability to resist it, how have we not bottomed out yet. Part 3 discusses this by nominating reasons for a deceleration of our downfall. Well if everything

seems rather bleak, what holds us from our incentives charging us rapidly downhill? "Why do things not degenerate..." Three basic reasons for the slowed, but inevitable, downfall. Excess Resources - we haven't reached the critical breaking point the Mathusalan rats experienced yet. Physical Limitations - there's literal physical limits to how far we can run downhill (eg. #of babies a woman can bear) Utility Maximization - "We've been thinking in terms of preserving values versus winning competitions, and expecting optimizing for the latter to destroy the former." However, fulfilling utilities sometimes need values to be optimized - although the equilibrium is fragile. eg. CSR to be a good firm. Greed doesn't bear capitalism, capitalism bears greed in people... Coordination - Although the lack of coordination is the main reason of these traps, subtle but potent coordination systems especially social codes are strong enough keeps us out of traps by "changing our incentives"

Tech Is An Accelerant

In this part, Alexander takes away the slight bit of hope that these 4 brakes introduce to slow your descent by introducing a new dimension, time. Additionally, Alexander points out at the acceleration of tech to fasten the blow on this dimension with glim dystopian futures where tech/ai eliminates each of the four brakes in Part. 3.

Well we'll reach these multipolar traps -- even if slow. Time is a relative, but key scale. Time is thus a dimension worth discussing. Time is further pushed by accelerated growth in technology. We can break the brakes in part 3, by reducing/removing physical limitations, for example. Tech deduces utility maximization as there is reduced need for human values, and coordination is unlocked to a new level by tech. Alexander further dramatizes the dimension of time and exasperation with technology by using a.i. dystopian futures. "The last value we have to sacrifice is being anything at all, having the lights on inside. With sufficient technology we will be "able" to give up even the final spark."

Once The Genie's Out The Box, There's No Going Back.

Gnon - nature, and its god - operates within Newton's third law of action necessitating a reaction. Gnon is basically Nick Land's version of Moloch. Violating these nature's laws through civilization leads to Gnon's wrath and our downfall. Gnon is a punishing god with no escape. **Reality Is Seemingly Sad**

The future is bleak, and Gnon is just another exemplification of Moloch. Submitting to them and following the "natural order of things" isn't going to make you "free". There is no order! It's always downfall. **Alternatives To Inevitable Downfall?**

So what now? Given that Moloch/Gnon or whatever wants us, and everything we value (i.e. art, science, love, philosophy, consciousness) dead, defeating them should be a high priority. Alluding to Bostrom's Superintelligence whereby the design of an intelligent machine will create a feedback loop of out-intellegenting itself. Given our action plan should be designing computers/intelligence that is smarter than us, but still keeps human values. But contrary to hubris where expecting god to wall us off if we submit to him, this Alexander proposes a transhumanist movement that is "rather actionable." Remove God from the picture entirely. As he puts it, "I am a transhumanist because I do not have enough hubris not to try to kill God."

- *Un-incentivized Incentivizer!**

Elua – the god of "... free love and all soft and fragile things" and mostly human values still exists. Even if the god seems weaker without worshippers, there he exists. As long as Moloch, the god where you can throw things you love to be granted power, exists, the offer is irresistible. A stronger god where we should help.

▼ Tulipmania and Taming Irrationality

My quest to understand market speculation —rather than relying on lazy quips about "animal spirits" or irrationality.



"Tulpenwoede" An Intro

Once in the Netherlands, tulips were worth more than real estate. Legend has it that in the 1630s, a sailor was thrown in a Dutch jail for eating a tulip, thinking it was an onion. At the time, the sailor's gluttony would have fed an entire crew.

Does an asset price rising to crazy heights with the help of ordinary investors hoping to avoid missing out sound familiar? It represents one of the earliest instances where asset prices deviate from intrinsic values.

During the period, wild speculation and euphoria of the masses led to "irrationally exuberant" spending on these bulbs. Ordinary citizens, even to the lowest dregs, were trading tulips. Properties were converted, assets to cash, and invested in flowers. Houses and land were sold at ruinously low payments at tulip markets. These tulips were loved for their deep, bright colors and exotic appeal and didn't experience price swings due to changes in production costs. Nor did they find new utility. Their popularity coincided with the Dutch Golden Age, where the republic was one of the world's leading economic powerhouses.

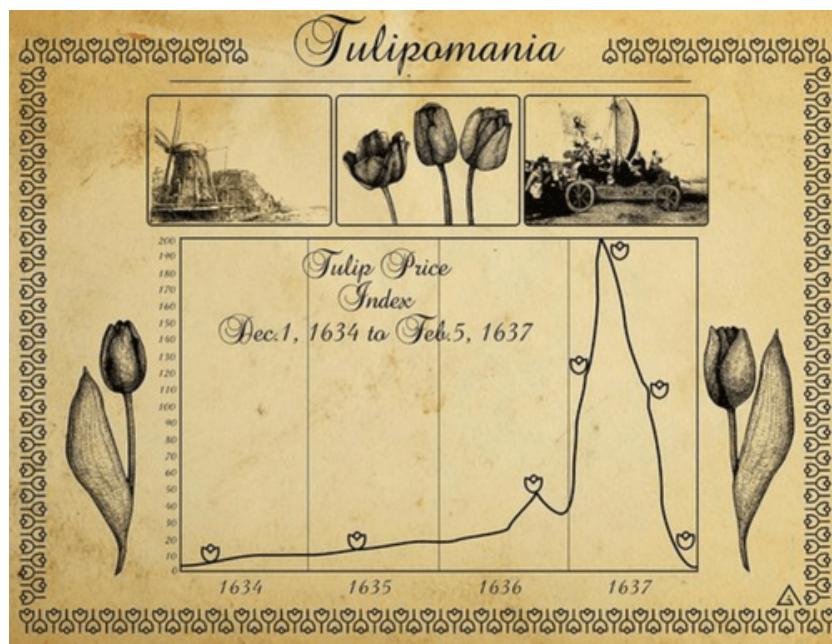
Per David Roos, post the 1620s depression, "...the Dutch enjoyed a period of unmatched wealth and prosperity. Newly independent from Spain, Dutch merchants grew rich on trade through the Dutch East India Company. With money to spend, art and exotica became fashionable collector items. That's how the Dutch became fascinated with rare "broken" tulips, bulbs that produced striped and speckled flowers."

During the event, historian Mike Dash mentions. Dutch artisans worked long hours for low wages. "When the day's work was done and they could finally go home, it was to cramped and sparsely furnished one or two-room houses that were in such short supply the rents were high...to people trapped in an existence such as this, the idea that one could earn a good living by planting bumps and sitting back to watch them grow must have been irresistible."

Post the rampage of The Bubonic Plague, there was a labor shortage, leading to higher wages and extra income for those who worked. Plus, the plague meant widespread lowered risk-aversion. The Dutch were fine indulging in speculation, knowing that each day could be their last. There was a post-plague "mood of fatalism and desperation," aiding speculation and reckless spending. The rich are accelerating prices even higher, buying rare breeds of tulips. Combining these factors, tulips increased in popularity as a means for people with disposable income to acquire wealth for the first time in many years.

Tulips were being sold for more than 10x the annual income of skilled artisans, and people kept on pouring life savings into buying tulip contracts anyway. When confidence was at its peak, everyone imagined their passion for tulips would last. I'd imagine those profiting trading bulbs could not resist telling family and friends of their good fortune.

There are stories of a man that sold his house in Hoorn Town for three bulbs — i.e. the first speculative bubble in history.



Calming Hand Taming Irrationality

Like all bubbles, in 1737, the market burst with groups of auctioneers lowering prices with no buyers. Due to the lack of interest, the market disappeared entirely in the coming days. However, investors acted rationally. According to Nicolaas Posthumus, a Dutch historian, serious tulip financiers generally did not participate in the speculative markets. The "mania" was usually self-contained within smaller circles and pushed by "casual traders".

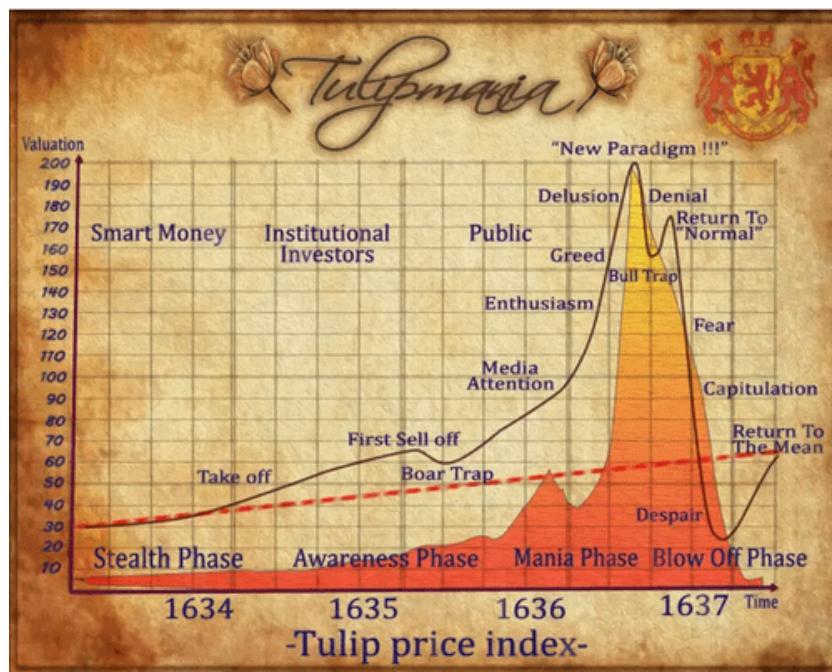
It is easy to claim that bubbles are irrational. They seem to represent a deviation of prices from fundamental values and contradict the basic economic theory. But there has been little attempt to understand the details of ***how speculation and the government are intertwined.***

Earl Thompson argues the market for tulips was an efficient response to the government conversion of futures contracts into options contracts. This was a deception by the government officials hoping to make a quick profit. The conversion meant investors who had bought the right to buy tulips in the future were no longer obliged to buy them. If the market price isn't up to one's liking, the investors had the option to pay a fine and get out. This increased tulip options prices, then collapsed when the government saw sense and canceled these contracts. The spot price and futures prices weren't volatile. Tulipmania was only a contractual artifact. Contrary to popular interpretations, there was no actual "mania."

The critical concept of preventing actual "mania" from happening is thus the government's calming hand. During the dutch times, corrupt officials realized their pursued ruse would cause mania, and stopping the conversion was a calming hand to their own doing. Nowadays, the government buoys speculators through unconventional monetary policies like quantitative easing — printing money to buy government bonds and mortgage securities.

The calming hand of governments nowadays is through unconventional monetary policies that are deemed to not encourage speculation — rather dampen it. However, economists worry that investors have come to rely on this calming hand of central banks. Unconventional monetary policy has been attacked for promoting further financial gaming. When it is taken away, speculative urges return. Central bankers might feel pleased with themselves for having tamed "animal spirits," but market uncertainty edges back in the weeks after monetary policy intervention.

This reliance has developed to a point where now, without regular interventions, markets become increasingly skittish. Central banks used quantitative easing and other monetary policies to save the world from financial meltdown. But easy money repressed, rather than extinguished, speculative practices. To feel comfortable halting these unconventional policies, central banks must ensure that the probabilities of nasty-tail risks have fallen. But can they ever do that? Hmm...?



▼ Expansion of Mobile Money in Ethiopia

Summary

I looked at Ethiopia's current business climate for mobile payment solutions that is financially inclusive. The current reach of mobile money has left the unbanked population that would've greatly benefited from the services, untouched. During the course of the paper, I look at regulatory reforms following 2018's government change and other regulatory reforms to establish ground for additive services between EthioTel and mobile money incumbents to include unbanked population. Then, I look at the economic viability, market size, and economic considerations of the symbiotic relationship established on the telecom giant's infrastructure. Finally, I look at how current sociocultural complexes could be navigated and benefit from the solutions suggested. \

Introduction

Saying Ethiopia's economy is cash-dominated would be an understatement. Only 31% of the population has bank accounts, making financial services in rural areas close to impossible. Borrowing money and other financial services take place through mediocre Micro-Financing Institutions (MFIs) and local savings clubs. (A) Mobile money is making a significant impact in bridging the digital divide between the developed and the developing countries, making millions of poor people use devices to transfer money, pay for goods, and access sophisticated financial services. (Dermish, 2007) The recent regulatory climate of the Abiy revolution facilitates the formation of mobile money services that don't require bank accounts. In this midst, partnership with EthioTel and incumbents would create great potential. Along with the right endorsement and orientation, it could reach unbanked regions, improve saving, and fuel growth in Ethiopia. 1 Regulatory Environment

Regulatory Environment

Since coming to power mid-2018, Ethiopia's Prime Minister Abiy has promised to "openup" the economy and loosen its monopoly on state-owned enterprises. (A) Ethiopia's highly regulated macroeconomic environment includes state ownership of the sole telecommunication provider – EthioTel. The commitment to liberalization started with partial privatization EthioTel and Ethiopian Airlines - Africa's biggest flagship carrier. (A)

PM Abiy's move also included an extensive overhaul of the financial sector. To boost noncash payments, the government announced the successful Kenyan mobile payment solution - MPesa would enter Ethiopia. (A) However, government doors were shut before completion of the deal. The sudden move was directed at excluding foreign fintech from reaping the business benefits and potential of the Ethiopian market. Plus, M-Pesa was considered to stifle local innovation. (A) Soon after, The House passed a bill September 2019 authorizing non-financial institutions, including EthioTel, to engage in mobile money services. The liberalization of EthioTel to private investors and newfound ability to participate in financial services allows partnerships with existing mobile money companies to emerge. A symbiotic relationship would morph the widescala network and userbase from EthioTel's side; with payment infrastructure, institutional bank relationship, and payment agents from MBirr/CBE-Birr's side. There are multiple advantages of employing existing local firms for mobile money solutions. One is the ability to prioritize unbanked regions, as urban regions are already within their userbase. Second, these firms participate in developing social values of saving and investing. All the while, transaction trend data could be used to inform policy decision making in the future.

Economic Considerations

Ethiopia's economy has been growing with double digits over the past 10 years and will continue to thrive in 2021-24. (A) It also has high levels of FDI that will incentivize the government forward with similar reformist agendas. However, operational mobile payment platforms have had limited growth. All service providers have no banking license, which allows them to provide the service directly to customers – essential for unbanked citizens. So, platforms have been targeting banked, urban users that saw limited utility. M-Birr, Ethiopia's first mobile money based on two banks and state microfinancing firms, only has 1.2 million users. Similarly, CBE-Birr (affiliated with Commercial Bank of Ethiopia), Hello Cash (from Cooperative Bank of Oromia) (A), and Amole (operated by Awash Bank)(A) have had hampered growth. (A)

Amidst all of this, EthioTel has been growing tremendously over the past 7 years, reaching 44% of the population, while smartphone internet penetration lags. EthioTel's widely available SMS SIM will have a hand in deriving better reach and inclusion of financial services – even without internet access. However, conflicts of interest will arise if Ethio-Tel decides to proceed with mobile payment services on its own – even after partial privatization. Instead, Ethio-Tel's partnership should provide the SMS infrastructure needed to support M-Birr, CBE-Birr, and others in providing financial inclusion to non-banked. Mobile money has the potential to reach unbanked people with phones, most of whom are under the government safety net. Ethiopia's Ministry of Finance could see significantly better efficacy from delivering Productive Safety Net

Program's financial assistance through mobile payment - contrary to cash where funds often get embezzled. The bill passed also requires a minimum of 50mill Birr and at least 10 shareholders to apply for a mobile money service license. This hurdle makes entrants more trustworthy, accountable, and sizable enough for healthy competition. Thus, strong capital markets and venture money going into mobile money, which has proven a lucrative investment in other African nations, will be of great benefit. To compete with more prominent incumbents (M-Birr and CBE-Birr), startups should form coalitions and agree with banks. That would strengthen their reach and potential to support transactions backed with assets.

Social and Cultural Considerations

Ethiopians are generally skeptical of innovation. They have a hard time trusting newer institutions, and legacy ones prevail – even with sub-par offerings. A past survey done in rural banked communities indicates that most people would rather walk an average of 3-4 miles for bank locations to find that ATMs are non-functional than use mobile payment methods. Mistrust emanates from thinking mobile money is independent of government control. Thus, endorsement from financial institutions, backing from EthioTel, and advocation from government bodies goes a long way in assuring communities.

Plus, Ethiopian's are recognized for their short-term-orientation, especially in rural areas. The saying "Worrying doesn't take away tomorrow's troubles; it takes away from today's peace" is usually taken out of context to oppose saving culture. Lack of financial inclusion doesn't help. The government's repeated trials to improve saving could benefit from mobile money solutions. Past studies on other African countries with mobile money solutions have shown an improvement in the likelihood of saving by 10.9%. (A)

Finally, entrepreneurship has been growing over the past five years due to increased backing from the government, a high number of STEM graduates, and jobless rates going up. Technological innovation has been on a steep climb, building Addis Ababa's Sheba Valley. However, a major impediment in the new ecosystem is the lack of payment gateways that support audiences these startups are targeting. Current API's don't support the non-banking population, significantly limiting the market size and ability to develop economies of scale. The start of this service would spur growth in companies that offer online services, including e-commerce and delivery, fueling growth.

Pending

▼ Modern Data Stack this Modern Data Stack that...

We've arrived at a point where the data landscape is a maze of tools, each serving a very specific purpose but often leading to a tangled web of integrations.

- The result? An overwhelming number of back-office processes that need to be managed, maintained, and understood just to keep things running.

In traditional data workflows, data cleanup and structuring often happen as a back-office process—an expensive, time-consuming endeavor that demands constant attention.

- But what if we could flip the script? What if the messy, unstructured data could be cleaned, transformed, and structured the moment it enters your system, right at the edge?

Instead of building a complex ecosystem of tools that need constant upkeep, what if we **frontloaded more of these processes directly into our applications?**

- By simplifying the architecture and placing the emphasis on front-loaded processes, we can create a more direct path from data to decision-making—without the detour through a dozen different platforms.

What if, rather than relying on a mess of back-office data tools, we designed our systems to handle data transformation and integration **closer to the user-facing side** of things?

We all should rethink our approach and bring data processes closer to the application layer, we can cut through the clutter and complexity of the data tool market

▼ Pirated Economist WebPage

▼ My Meta post-mortem on working with “wild-data”

- We looked at two main metrics for every cut - YoY and Variance from Goal.
 - YoY fraction has two components that tell you is it last years performance or current year performance that's affecting change
 - Variance fraction has two components that tell you is it performance or goal that's the issue