I'm working on web3/DeFi-native modelling and am seeking funding + colab opportunities for the next stage. Wanted to gauge community mood regarding match of interests before submitting formal applications.

TL;DR A web3 protocol is an evolving shared narrative: it lives simultaneously in two domains open for exploration, onchain transactions and community interactions. They inform and impact each other in both directions in a myriad of ways creating a bipartite heterogenous living organism. To comprehend and be able to predict this organism means to be able to model and predict evolution of its two integral parts in a coherent, unified way. Currently available intelligence tools don't leverage web3 nature, but merely translate approaches from web2/TradFi, hence falling short of providing deep insight. This proposal strives to make the first crucial step towards solving this problem.

web3/DeFi intelligence at the moment is at theskeuomorphic stage — web2/TradFi tools like SQL data queries, basic accounting, node ranking, outdated VaR risk modelling, agent-based modelling backed by the rational choice theory are translated as if there's no fundamental difference between TradFi, where part of data is simply not digitised, and the bulk of digital data is private, and web3, which leaves opportunities for insight provided by the latter unexplored.

Expanding on a16z' Chris Dixon we can postulate that a web3 protocol with its products is an evolving shared narrative: be it a token, NFT, web3 games, decentralized social media, or any yet to be invented thing, its utility and value is a function of a changing commonly shared narrative, where the essence of the narrative, the nature of changes and commonality play equally important parts.

A web3 protocol lives simultaneously in two domains open for exploration: onchain transactions and community interactions. They inform and impact each other in both directions in a myriad of ways creating a bipartite heterogenous living organism. Loop: people engage in discussions, track discussions, track onchain data, make decisions, execute transactions, other people track these decisions and rationalisations and make their own decisions. To know and comprehend this organism means to be able to model and predict evolution of its two integral parts in a coherent, unified way. Luckily, in DeFi both financial and contextual data, transactions and public sentiment about them on Twitter/Discord/Discourse, is fully open and available for modelling and insight.

Hence, in DeFi/web3 we can build native decision making around native, non-skeuomorphic intelligence. Specifically, two major opportunities here are to learn models from real time data of public sentiment towards specific protocols/products and learn models from onchain data with vault-level granularity. And since they impact each other supermodels uniting both forces can be built as well.

### UST sentiment => crash case

A few words about opportunities around sentiment. Why sentiment matters? The recent UST crash was the result of a gradual crowd sentiment deterioration => loss of confidence => panic => bank run => death spiral supported by the mechanics of the protocol. We just witnessed how public sentiment dynamics wiped off $40 bln of value over several days.

[

](https://aws1.discourse-cdn.com/standard20/uploads/gro/original/1X/eb7bd53f60ff210f36a7ef24e8d7860ee09045c2.png)

Twitter scraping code || Raw tweet data

Language model: BERT_base model (12-layer, 768-hidden, 12-heads, 110M parameters), RoBERTa pretraining procedure. Pretrained on the corpus of 850M English Tweets (16B word tokens ~ 80GB) streamed from 01/2012 to 08/2019. Fine-tuned for sentiment analysis with the SemEval-2017: Task 4A dataset.

Sentiment analysis code || Sentiment data: result of sentiment analysis || Visualise the result

Rest assured it's not an isolated case. Already in 2010it was shown that the accuracy of a model predicting daily closing values of Dow Jones Industrial Average is significantly improved by the inclusion of specific public mood dimensions as measured from Twitter feed.

Moreover, realising the importance of public mood TradFi public institutions, includingFederal Reserve and World Bank, are tracking public sentiment.

Back in 2010 sentiment was measured with a fixed lexicon, now we can use large language models fine-tuned with domain-specific annotated datasets, which would provide precision insight into events like. UST crash.

Models learned from social networks & onchain time series can also be used in asset management and broader DAO decision making mapping public mood and onchain dynamics to (future) TVL, usage or token price, discovering patterns of challenges and opportunities, running simulations, optimising vault design and parameter set. Among many other things.

For instance, UST crash provides an excellent dataset of matching sentiment and onchain dynamics of real-time DeFi-native panic pattern, which could be learned into a model and used to discover future similar cases.

## PROPOSAL

The big opportunity here is to design, build and train a model, which intakes community and onchain data relevant to the protocol/product and outputs respective evolving shared narrative. Then build decision support tools around this web3-native intelligence.

[

](https://aws1.discourse-cdn.com/standard20/uploads/gro/original/1X/b433e7a3321d2baff81b05504111f6be86439cce.png)

I demonstrated above a proof of concept. To move it towards production progress should go in a number of parallel ways.

The first stage — three-four months — of this work will encompass the following :

## I. Build a custom DeFi-focused annotated community interactions dataset to fine-tune a large language model for multidimensional affect analysis

To gauge public sentiment I use a general purpose language model pretrained on the large corpus of texts and then fine-tuned with the SemEval-2017: Task 4A dataset for sentiment analysis.

The result is impressive, but we can get far deeper and richer insight.

Currently state-of-the-art natural language processing results are achieved with Transformer-based large language models (OpenAI's GPT-3 — 175B, DeepMind's Gopher — 280B params). Until most recently such models were accessible only to the BigTech, but the Big Science project is changing the game: an international non-profit consortium is working on the SoA multilingual (46 langs) BLOOM LLM, 176B params, pretrained on 350B words, which will be fully open from day one. Pretraining started on May, 11 and will finish within 3-4 months. It's our chance to ride the wave.

To leverage the full potential of this opportunity we need to use coming months to get prepared: build a custom DeFi-native labelled community interactions dataset to fine-tune BLOOM LLM for our purposes. DeFi conversation on Twitter/Discord/Discourse is full of domain-specific slang, nuances, deep context and a general purpose model must understand these details natively to output the most adequate results. Moreover, the SemEval-2017-4A dataset, although popular, contains 50k tweets and only 3 classes of sentiment (positive/neutral/negative). The bigger dataset for fine-tuning is the better results our model will produce.

So we'll use the most recent DynaSent sentiment analysis benchmark dataset as a size reference (130k). Also, clearly affect expressed by people is far more nuanced than what could be captured by a three-class scale, which will be reflected in a bigger, more nuanced scale.

[

](https://aws1.discourse-cdn.com/standard20/uploads/gro/original/1X/21fdb3066d370809f10a3bfee8e83272dff3a861.jpeg)

Although there are multiple commercial companies offering data labelling services,it's been shown that retail annotated datasets contain systemic labelling errors. This is especially urgent for us given that we're looking into identifying affect of a tweet/post. Hence, we need to assemble a group of handpicked amply qualified annotators, train them to properly understand DeFi landscape and instruct them to annotate a dataset assembled from Twitter/Discord/Discourse discussions with labels reflecting mood of each tweet or post. We'll be using Heartex to run bespoke labelling.

Nobody has ever assembled such dataset for DeFi; if performed, this asset will give us a huge edge; this investment will bring immense return down the road.

Future versions of the community mood model could be enriched with a community social graph data. Pure language models presume that all social interactions influence equally the state of community mood, which is clearly not the case. A message from a protocol founder vs an average troll clearly should have different weight in the resulting picture.

One way to approach this is to use a Temporal Graph Network (TGN) to represent an evolving community graph. TGN generates temporal embedding (a real-valued vector) for each user (graph node) i. This embedding is a learnable function of their history of interactions (messages) and that of their n-hop neighbours (social ties akin to PageRank).

## II. DL experiments on the second component of the Big model, the vault-level model of onchain transactions.

A Spatio-Temporal Graph Network (STGN) can be used to represent an evolving graph of onchain transactions. Here, say for a given vault, STGN generates a temporal embedding (a real-valued vector) for each wallet (graph node) i. This embedding is a learnable function of their history of transactions (messages) and that of their n-hop neighbours:

[

](https://d2kq0urxkarztv.cloudfront.net/6130cc83941c6200a154ea23/3104296/upload-17923e5b-6bc8-478e-a72e-2da2cd2d6091.png?cX=0&cY=1&cW=1150&cH=1369)

A network (ETH)-wide representation graph can also be built.

A vault model can be used for prediction: running test scenarios for different assets/platforms/DAO decision parameter sets; classification: risk/value profiling assets/platforms. Tx patterns forecasting. Fraudulent activity detection. Detecting other patterns/clusters on the trie providing certain insights and serving as a decision support tool for all stakeholders: DAOs, investors, community members, users. I invite everybody to brainstorm further possible applications.

Again, to the best of my knowledge nobody has ever done this before, this is an investment with huge potential return.

The most interesting and complicated part further down the road is to merge two components of the bipartite web3 model.

One approach: both models can be regarded as functions, each in their respective function space. Then temporal evolution of these models can be modeled as an operator on this function space; then a learnable mapping between these function spaces will be a representation of mutual impact of community life and onchain transactions bringing unity we're seeking here.

Another approach: use Bayesian Networks to identify mutual causation.

I've studied under Prof. Michael Bronstein, a specialist in geometric DL and Head of Graph Learning Research at Twitter, and I plan to reach out to him regarding this project; his expertise would be particularly helpful here.

## III. Finally, build a UX/UI design concept to get adequate insight into the data.

All data and models in the world are worthless in the absence of adequate HCI tools. We must be able to dive into the evolving narrative about community mood and its bijective mapping to the onchain dynamics in an effortless, yet profound way.

For this part I plan to bring in the art director I know personally. He is responsible, among many other projects, for the Offshore pipeline management system interface for South Stream (iF Design Award '22; Red Dot Design Award '21)

[

](https://aws1.discourse-cdn.com/standard20/uploads/gro/original/1X/dac677faf4d01529ed1f261730d1abf9c0a92be0.jpeg)

[

](https://aws1.discourse-cdn.com/standard20/uploads/gro/original/1X/58c5c0b449943242fb39040d6dd8b6072afc1357.jpeg)

and Moscow Metro map 3.0 (iF Design Award '17)

[

](https://aws1.discourse-cdn.com/standard20/uploads/gro/original/1X/d792b0e12299772fde1a5388b9e1cc40544d7bff.jpeg)

The purpose of this part of the proposal is to create a concept of a design system/design language, which will later be leveraged to build an interface into the shared evolving narrative. The concept will be built in Unreal Engine, it will be a live simulation aka infinite game-inspired tool for diving into the evolving shared narrative == a web3 product.

[

](https://aws1.discourse-cdn.com/standard20/uploads/gro/original/1X/a063ad21dba381fc0619edd6b68b15c5e3ab76f1.jpeg)

It will also hopefully bring some good publicity once built.

A recently published DeepMind's Gato model represents an early step towards what I'm implying as a horizon for this proposal: a multi-modal, multi-task model, which intakes heterogenous data, like social network discussions and onchain data time series, and outputs an integrative domain-specific prediction like a shared narrative regarding a certain topic over a certain period.