



On Decentralizing AI Data & Services

Trent McConaghy
@trentmc0

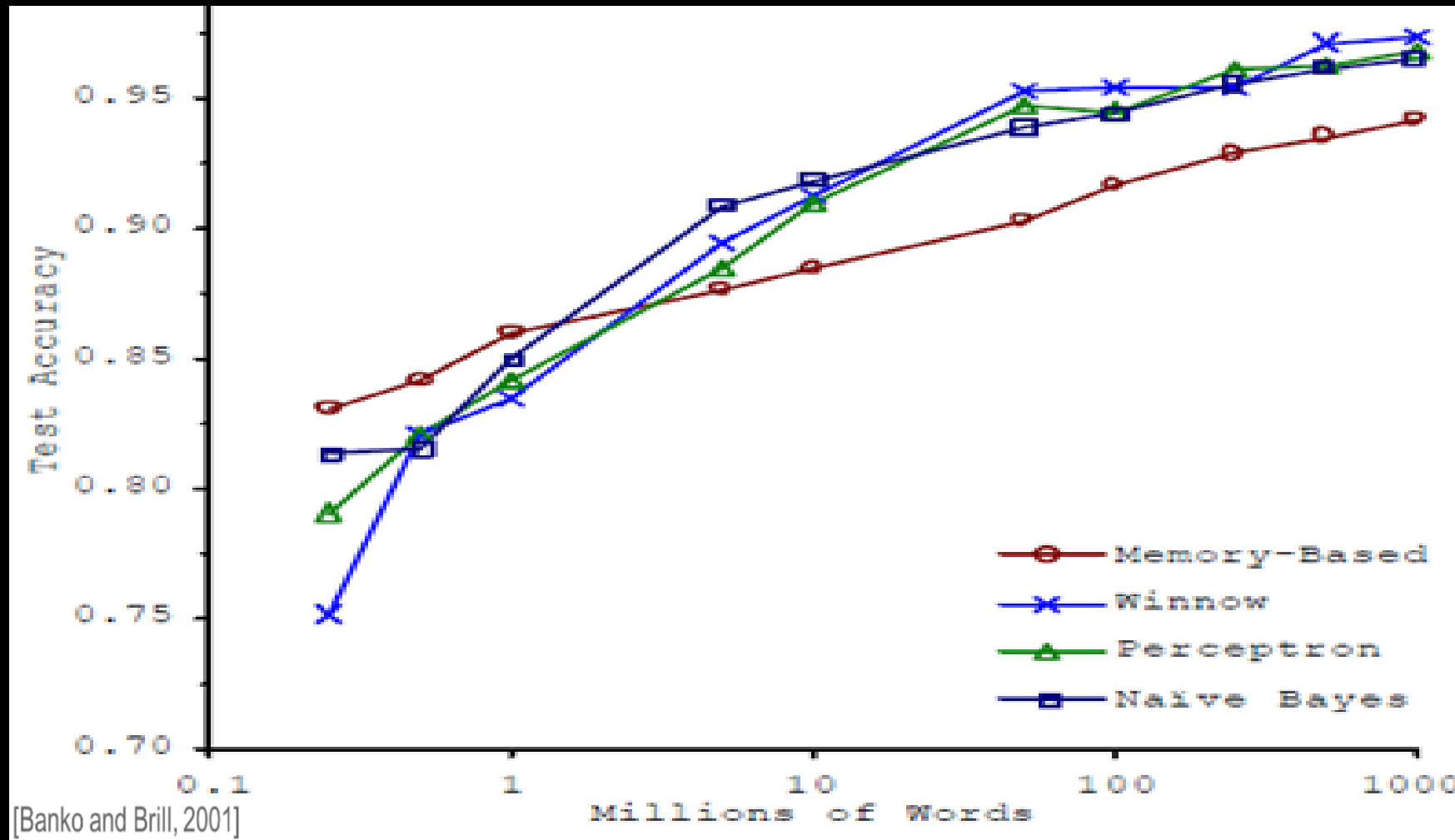
#Data
#Incentives





Audio radar

The Unreasonable Effectiveness of Data



1000x more data

1000%
less
error!

The world's most valuable resource



Data and the new rules
of competition

Silo ++ data



++ accuracy



++ \$

Default incentive:
hoard the data

**“Show me the incentive
and I will show you the outcome.”**

-Charlie Munger

You can get people to do stuff
by rewarding them with tokens.
This is a superpower.





Change the
incentives!

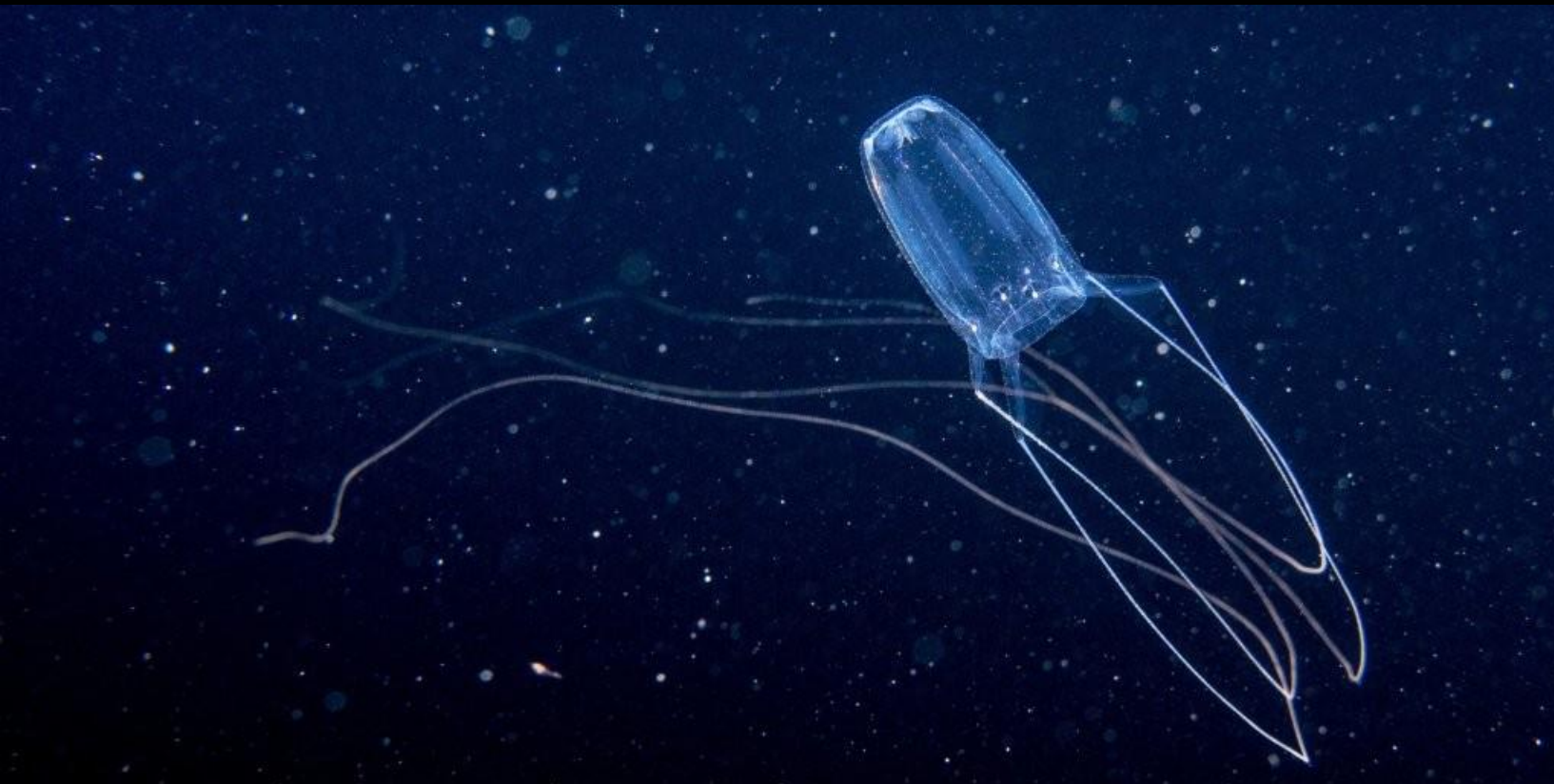
~~Site~~ *Pool* ++ data



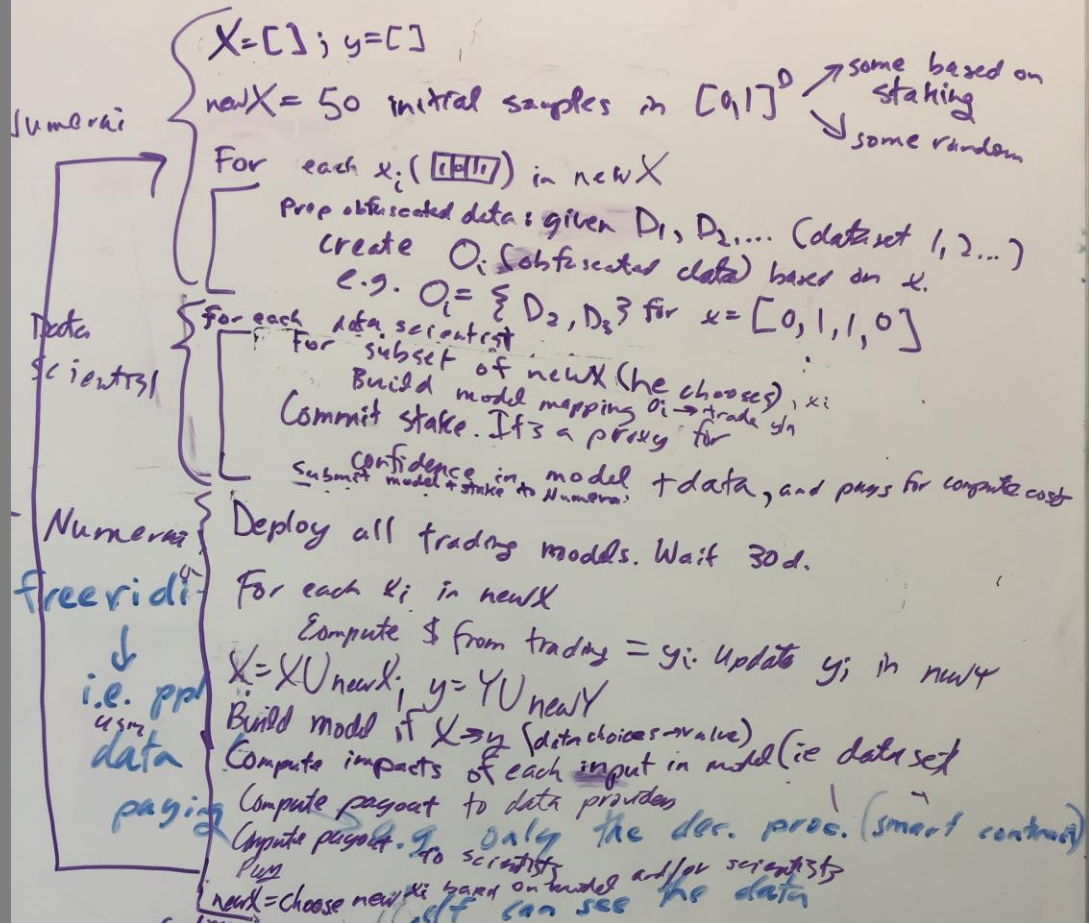
++ accuracy



++ \$



**Early
iterations**

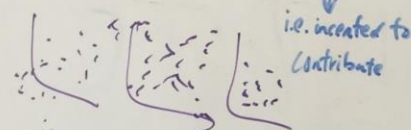


- homomorphic encr.
- 2kp to trustless computing
- intd/ms secure enclaves
- latent variable obfuscation / GANs

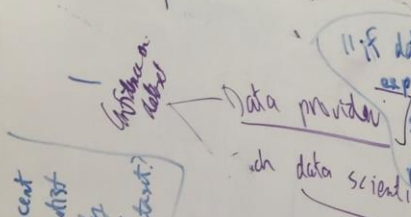
Goals:

1. Healthy ecosystem ^{initially} \ long term
2. Maintain ethics/values $\rightarrow$ governance

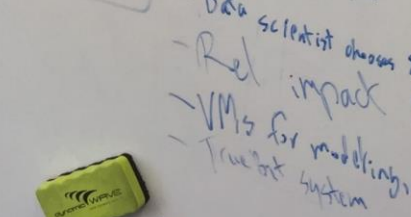
- Data is open (but obfuscated)
- Avoid need for TTP to see data when Company
- How to get people compensated w/o freeriding



- How to price data
 - fungible
 - non-fungible

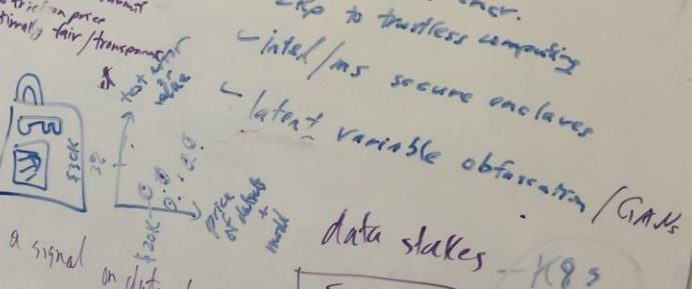


- Feature eng on plaintext
- Data scientist predicts rel. impact
- Data scientist chooses % of returns to all data to each dataset
- Rel impact
- VMs for modeling w/ permissioned just test
- Trust system

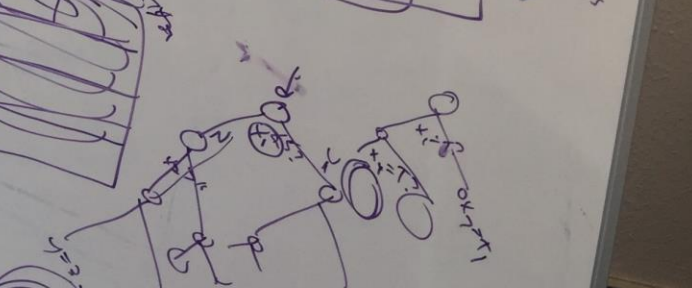


- MLM on datasets
- Scarcity for use of dataset. Data scientists bid on stake to use.

for subset of new X (he chooses) x_i
 Build model mapping $0 \rightarrow 1$ trade y_i
 Commit stake. It's a proxy for confidence in model + data, and pays for com of model
 Submit model stake to Numeras
 Deploy all trading models. Wait 30d.
 For each k_i in new X
 Compute S from today = y_i . Update y_i in new X
 $X = XU_{new}$, $y = YU_{new}$
 Build model $if X \rightarrow y$ (data choices - value)
 Compute impacts of each input in model (ie dataset)
 Compute payout to data providers
 Compute payout to scientists
 Only the dec. price. (Smart contract can see the data) → New test



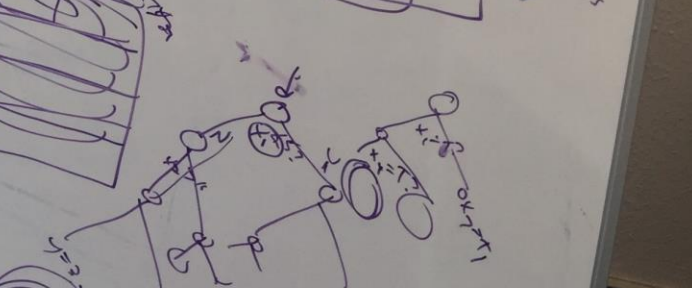
homomorphic enc.
 2k to trustless computing
 intel/ms secure enclaves
 latent variable obfuscation / GANs



data stakes

50000
10000
10
5

system for privacy



Multi-level auction, ^{user} set price

- I have
1. ~~\$20K~~ 1K miles, bidders please
 2. Auction happens

highest bid →

2nd-highest bid →

pre-set conditions based on top bid, #2 bid, ...

when you get to see if

\$10K	now
\$6K	in 1 mo
\$3K	in 2 mo
\$1K	in 3 mo

TOP 3 bids get data now, rest in 1-6 mos"

And: in 6 mos: data is set free.

Marketplace for obfuscating data

"Obfuscated data is only usable by Numerical ite(P)"

- Steamit / MCM

- Content's data
- post content, - post data, get tokens
 - send token to upvote
 - if others upvote, you get tokens

Pooling with

~~that~~ set price
Supplier-set price

1. Supplier says: total price \$20K. Top bidders in pool get data now.
2. Rest get data later.

\$20K set by supplier

10K	
7K	
\$22K	
1K 2K	

these folks get data now

others get data in 1 mo or 6 mos (data set free)

graduated DT's 1 mo, 2 mos, ...

Engagements / Incentives

fix price up-front

u " u

+ reputation (subjective)

- 1. You have \$50K staked in a wallet
 - 1/10000 missing wallet
 - each key is a percentage hash of a row of data (HD wallet)
- 2. Alg to compute a key from data is public
- 3. $\therefore$ if any data is made public, then anyone can open my wallet & get \$50K

markets

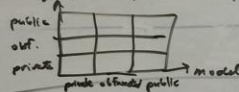
data $\rightarrow$ distance

no info $\rightarrow$ test error $\rightarrow$ diff



Stake

model params visible for data supplier
data visible for scientist
data private - model params private

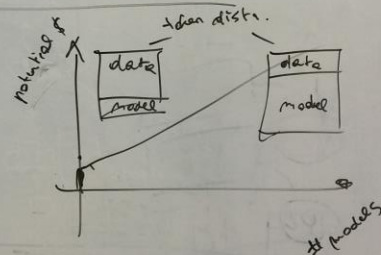


co-owner of \$

or Home Markets?

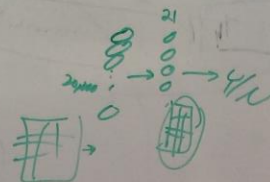
Tokenizing Access to Data Revenue, Fixed Supply

- Each dataset has its own tokens. Fixed supply
- "ICO a dataset"
- When data is purchased for use, \$ is split according to token ownership



Tokenizing Access to Data Itself

- e.g. 100 tokens. You can access the data if you own the token.



Challenges

0. Generate data.

- How to ensure supplier gets paid w/o losing ability to get paid in future. "Free riding"
"Privacy"
"Copy vs title"

- How to price?
Friction in pricing) — overall price
— ~~relative~~ relative impact per dataset

Static $\longleftrightarrow$ Dynamic dataset (goes stale)

Signals

- data labeling service
- data obfuscation market
- Stake in belief of ~~data~~ value of dataset $\leftarrow$ ~~supplier~~ ~~scientist~~ $\leftarrow$ by data scientist
- price asked by supplier
- price bid by scientist
- reputation of dataset $\leftarrow$ ~~supplier~~ $\leftarrow$ scientist
- ~~value~~ value gained from dataset(s) in ~~Real~~ RL (e.g. Namecoin & smart; insurance saving)
- total value of network $\leftarrow$ value per dataset
- prediction market belief in value of dataset
- novelty of a dataset
- taken price of data set
- ~~specificity~~ ~~element~~

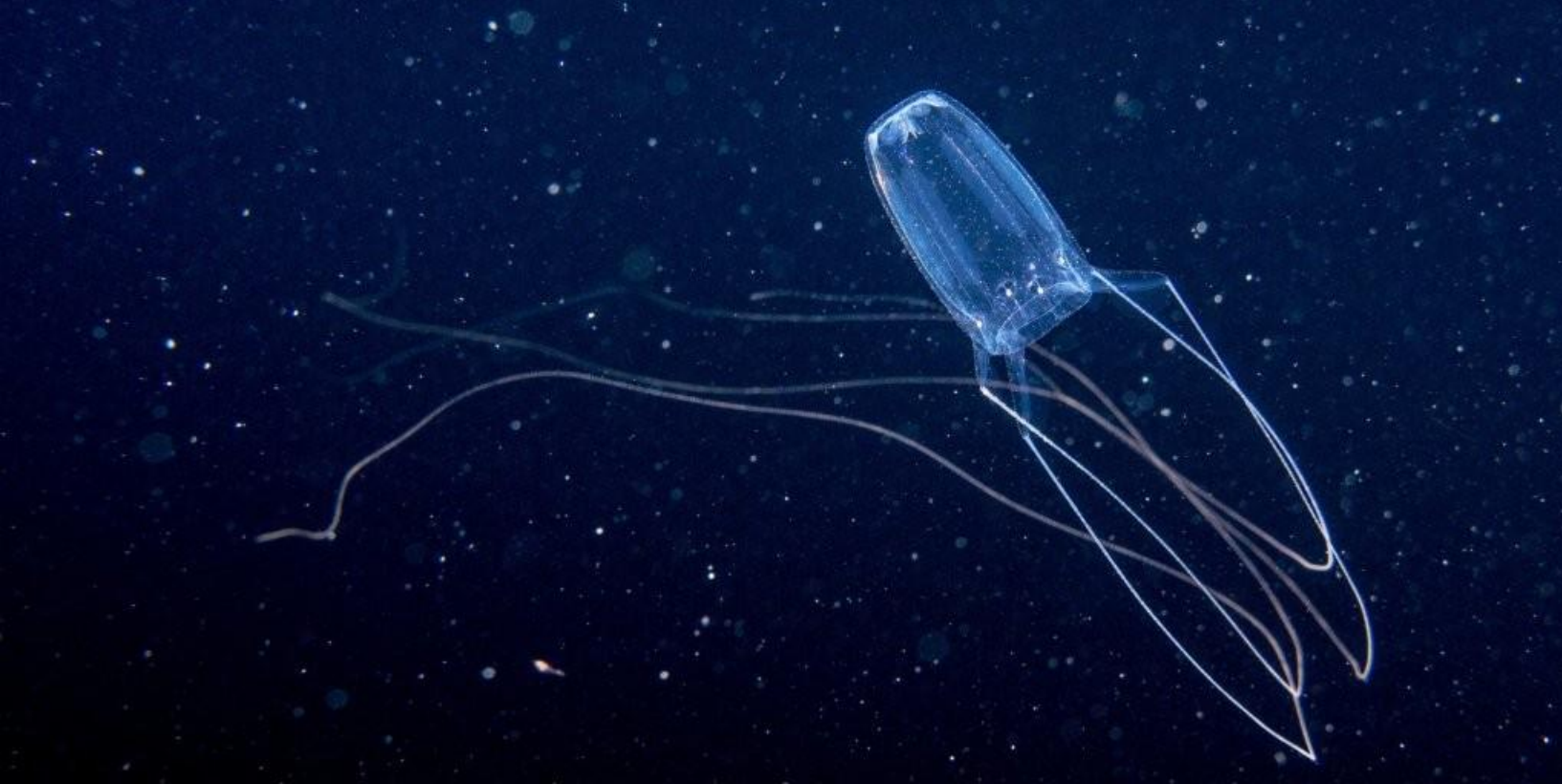
Tools to

- address free riding
- "Set the free" after $\Delta t \rightarrow$ stream to own
- Data provider doesn't care
- Licensing
- ~~free~~ ~~open~~ ~~source~~ ~~code~~ $\rightarrow$ combine
- watermarking
- reputation
- provenance
- risk of litigation
- only the smart contract can see the data. Eg doctor + locks $\rightarrow$ MPC, $\rightarrow$ RL
- If data set free, ~~you~~ ~~there~~ ~~you~~ private key gets exposed $\rightarrow$ end stake of \$\$\$
- Data obfuscation, e.g. latent variables on NN (like Normani)

Fungible $\longleftrightarrow$ Non fungible

Private via $\rightarrow$ ~~hyper~~ ~~market~~ ~~place~~ $\rightarrow$ ~~or~~ ~~Bitcoin~~ ~~pooling~~

Continued...
Estimates the value of the data



Early iterations:
Flailing

**Can we
structure this
better?**

Realization: Tokenized Ecosystems Are a Lot Like Evolutionary Algorithms (EAs)!

What	Tokenized ecosystem	Evolutionary Algorithm
Goals	Block reward function E.g. “Maximize hash rate”	Objective function E.g. “Minimize error”
Measurement & test	Proof E.g. “Proof of Work”	Evaluate fitness E.g. “Simulate circuit”
System agents	Miners & token holders (humans) In a network	Individuals (computer agents) In a population
System clock	Block reward interval	Generation
Incentives & Disincentives	You can’t control human, Just reward: give tokens And punish: slash stake	You can’t control individual, Just reward: reproduce And punish: kill

**We can approach token design
as EA design.**

EA Design



Steps in EA (and Optimization) Design

1. **Formulate the problem.** Objectives, constraints, design space.
2. **Try an existing solver.** If needed, try different problem formulations or solvers.
3. **Design new solver?**

1. Formulation of an optimization problem

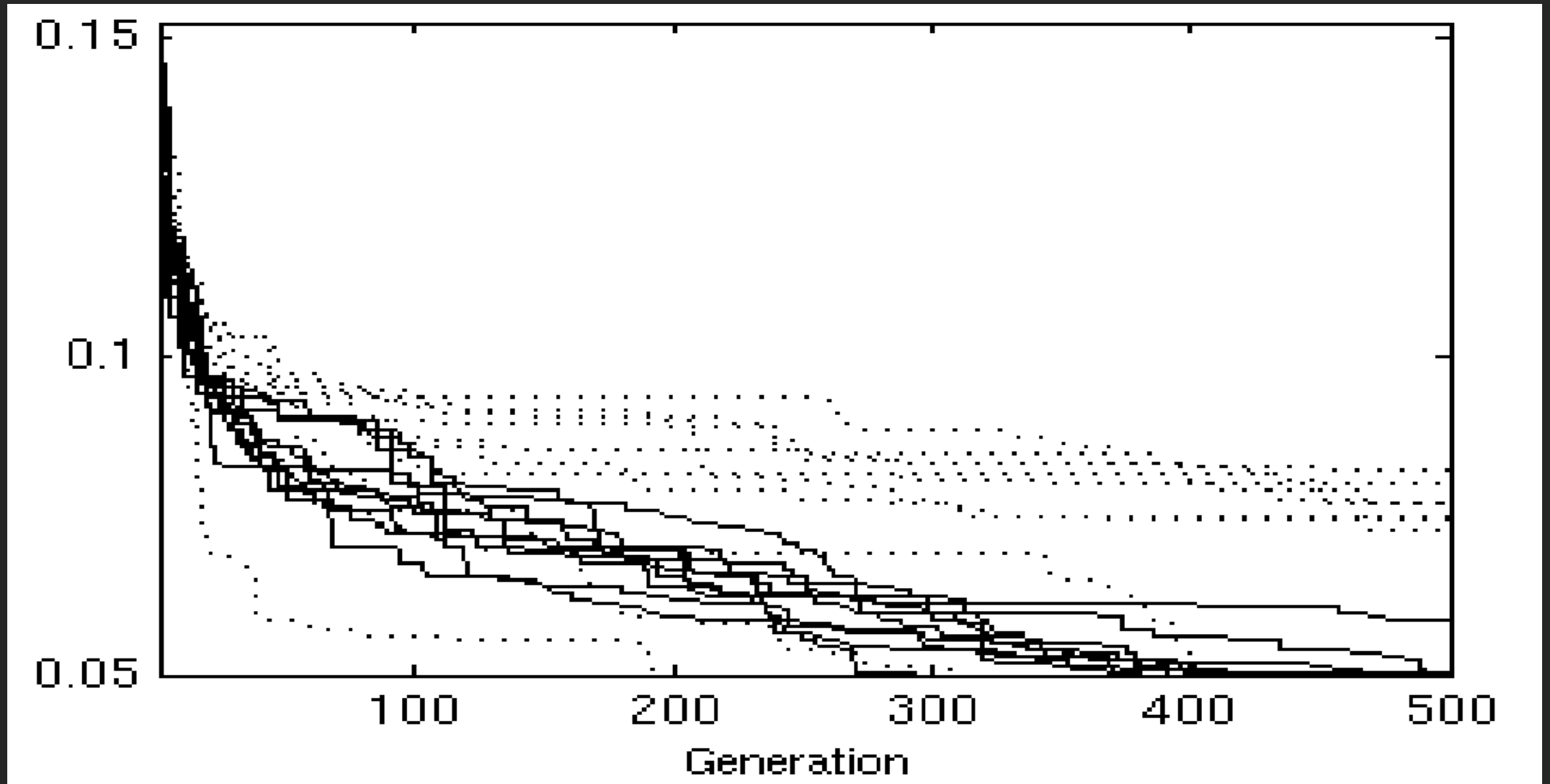
Objectives & constraints in a design space

The algorithm's aim is formulated as a constrained multi-objective optimization problem

$$\begin{aligned} &\text{minimize} && f_i(\phi) && i = 1 \dots N_f \\ &\text{s.t.} && g_j(\phi) \leq 0 && j = 1 \dots N_g \\ &&& h_k(\phi) = 0 && k = 1 \dots N_h \\ &&& \phi \in \Phi \end{aligned} \tag{1}$$

where Φ is the “general” space of possible topologies and sizings. The algorithm traverses Φ to return a Pareto-optimal

2. Try an existing solver. Does it converge?



3. Design new solver

TABLE II
PROCEDURE SANGRIAOPTIMIZATION()

Inputs: $D, N_a, K, N_L(k)$

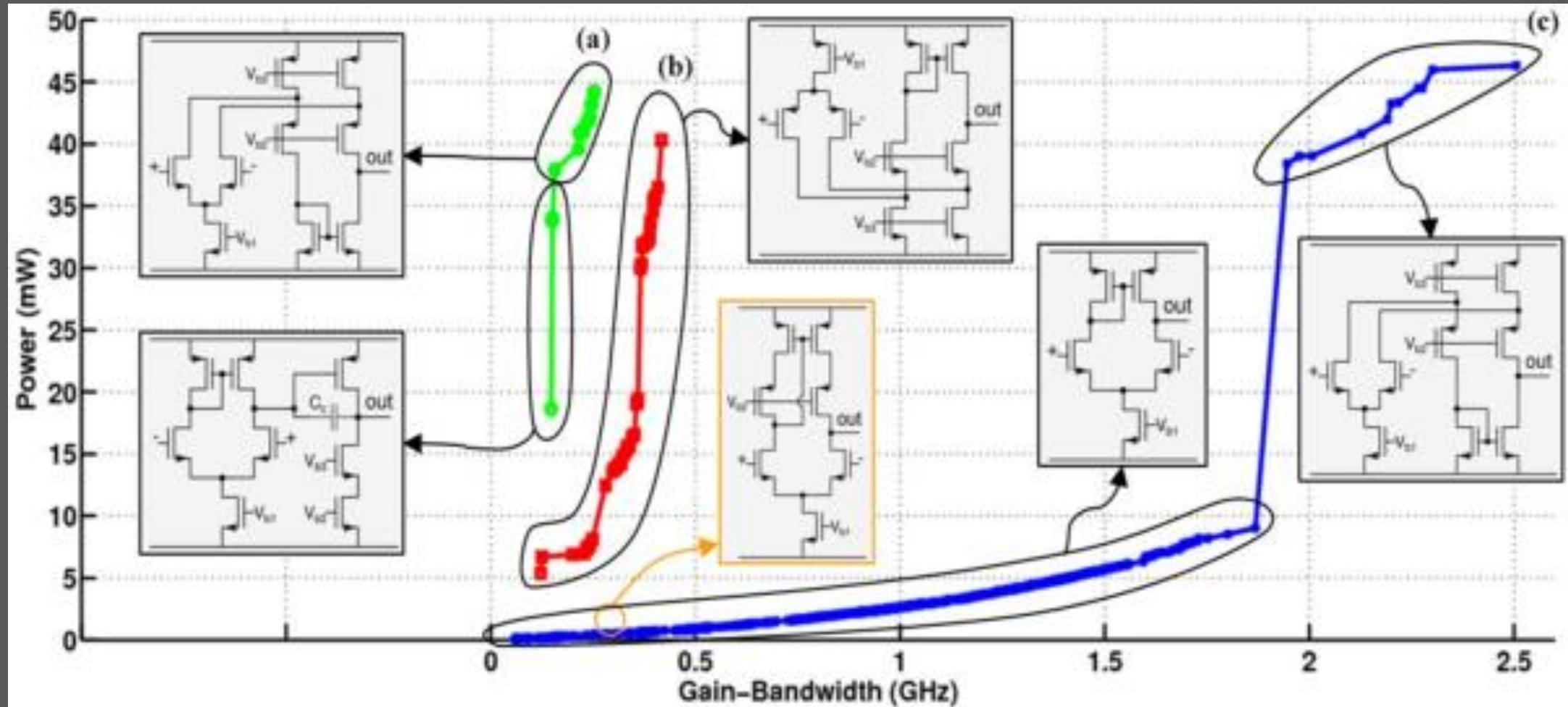
Outputs: d^*

1. $N_{gen} = 0; P = \emptyset, P_{all} = \emptyset$
2. while $stop() \neq True$:
3. if $(N_{gen} \% N_a) = 0$:
4. if $|P| < K$:
5. $P_{|P|+1} = \emptyset$
6. $P_0 = \text{SpaceFillIndividuals}(N_L(k), N_D, D)$
7. for $k = 1$ to $|P|$:
8. $P_k = \text{SelectParents}(P_k, P_{k-1}, N_L(k))$
9. $P_{k,j} = \text{UpdateLocalOptState}(P_{k,j}, k), j = 1$ to $|P_k|$
10. $P_{all} = \text{unique}(P_{all} \cup P)$
11. $P_{|P|} = P_{|P|} \cup \text{InnerOptimize}(P_{all}, D, k)$
12. $d^* = d_i$ in P_{all} with highest Y or Cpk
13. $N_{gen} = N_{gen} + 1$
14. return d^*

and all individuals encountered so far in the search, P_{all} .

Lines 2–13 are the generational loop, which repeats until stop

Example of a Successful Outcome



Token Design as EA Design



Steps in *Token* Design

1. **Formulate the problem.** Objectives, constraints, design space.
2. **Try an existing pattern.** If needed, try different formulations or patterns.
3. **Design new pattern?**

1. Formulate the Problem

(a) Ask

- Who are my potential **stakeholders**?
- And what do each of them **want**?
- What are possible **attack vectors**?

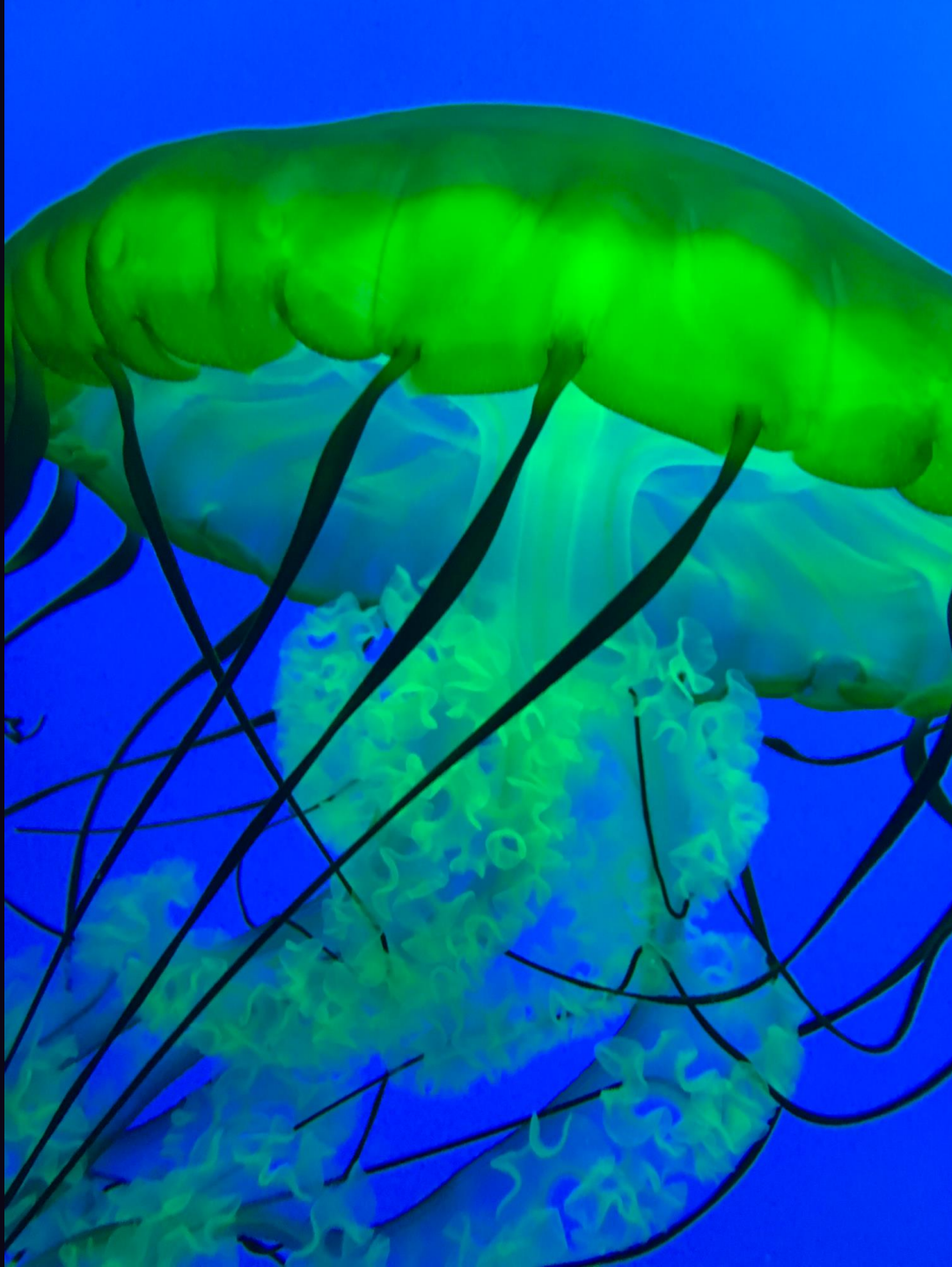
(b) Translate those into objectives and constraints.

2. Try Existing Patterns

1. Curation
2. **Proofs of human or compute work**
3. Identity
4. Reputation
5. Governance / software updates
6. Third-party arbitration
7. ...

2.1 Patterns for Curation

- **Binary** membership: Token Curated Registry (TCR)
- **Discrete-valued** membership: Stake Machines
- **Continuous-valued** membership: Curation Markets characterized by bonding curve
- **Hierarchical** membership: each label gets a TCR
- **Work** tied to membership: Proofed Curation Market
- **Non-fungible** tokens: Re-Fungible Tokens



Case Study: Analysis of Bitcoin



Bitcoin objective function

Objective: Maximize security of network

- Where “security” = compute power
- Therefore, super expensive to roll back changes to the transaction log



Bitcoin objective function

Objective: Maximize security of network

- Where “security” = compute power
- Therefore, super expensive to roll back changes to the transaction log

$$E(R_i) \propto H_i * T$$

$E()$ = expected value **block rewards** **hash power of actor = contribution to “security”** **# tokens (BTC) dispensed each block**

Result of Bitcoin's objective function:

People are maximizing security! = Maximizing electricity





Case Study: Design of Ocean

1. Formulate the Problem:

(a) Who are stakeholders? What do they want?

Key stakeholders in Ocean ecosystem

Stakeholder	What value they can provide	What they might get in return
Data/service provider, data custodian, data owner	Data/service (market's supply)	Tokens for making available / providing service
Data/service referrers, curators. Includes exchanges and other application-layer providers.	Data/service (via a provider etc), curation	Tokens for curating
Data/service verifier. Includes resolution of linked proofs on other chains	Data/service (via a provider etc), verification	Tokens for verification
Data/service consumer	Tokens	Data/service (market's demand)
Keepers	Correctly run nodes in network	Tokens for chainkeeping

1. Formulate the problem:

(b) Translate into objectives and constraints

Objective function: maximize supply of relevant data

Token rewards if: supply relevant data

Token rewards if: supply data, and curate it

1. Formulate the problem:

(b) Translate into objectives & constraints

Constraints = checklist:

- For priced data, is there incentive for supplying more? Referring?
- For priced data, good spam prevention?
- For free data, is there incentive for supplying more? Referring?
- For free data, good spam prevention?
- Does the token give higher marginal value to users of the network versus external investors? Eg Does return on capital increase as stake increases?
- Are people incentivized to run keepers?
- Is it simple? Is onboarding low-friction?

2. Try Existing Patterns

Some patterns:

1. Actor registry
2. Data registry
3. Actor registry + data registry
4. Data registry + free-as-in-beer data curation market.
Curation: Pay tokens to listen.

2. Try existing patterns: evaluate on objectives & constraints. None passed...

Key Question	1	2	3	4
For priced data: incentive for supplying more? Referring?	✗	≈	✓	≈
For priced data: good spam prevention?	≈	✓	✓	✓
For free data: incentive for supplying more? Referring?	✗	≈	✗	✓
For free data: good spam prevention?	≈	✓	≈	✓
Does token give higher marginal value to users of the network, vs external investors? Eg Does return on capital increase as stake increases?	✓	✓	✓	✓
Are people incentivized to run keepers?	≈	≈	✓	✓
It simple? Is onboarding low-friction? Where possible, do we use incentives/crypto rather than legal recourse?	✓	✓	≈	≈

3. Try **New** Patterns

Some patterns:

1. Actor registry
2. Data registry
3. Actor registry + data registry
4. Data registry + free-as-in-beer data curation market. Curation: Pay tokens to listen.
5. **Data registry + free data curation market. Curation: Stake tokens as belief in reputation. Auto CDN.**
6. **Actor registry + free&priced data curation market. Curation: Stake tokens as belief in reputation. Auto CDN. “Proofed Curation Market”**

3. Try **new** patterns: evaluate on objectives & constraints

Key Question	1	2	3	4	5	6
For priced data: incentive for supplying more? Referring?	✗	≈	✓	≈	≈	✓
For priced data: good spam prevention?	≈	✓	✓	✓	✓	✓
For free data: incentive for supplying more? Referring?	✗	≈	✗	✓	✓	✓
For free data: good spam prevention?	≈	✓	≈	✓	≈	✓
Does token give higher marginal value to users of the network, vs external investors? Eg Does return on capital increase as stake increases?	✓	✓	✓	✓	✓	✓
Are people incentivized to run keepers?	≈	≈	✓	✓	✓	✓
It simple? Is onboarding low-friction? Where possible, do we use incentives/crypto rather than legal recourse?	✓	✓	≈	≈	✓	✓

Objective: maximize supply of relevant data

- Reward curating data (staking on it) + making it available
- New pattern: Proofed Curation Market

$$E(R_{ij}) \propto \log_{10}(S_{ij}) * \log_{10}(D_j) * T * R_i$$

Expected
reward for user
 i on dataset j

S_{ij} = predicted popularity
= user's curation market
stake in dataset j

D_j = proofed popularity
= # times made dataset
available

tokens
during
interval

From AI data to AI *services*



Motivations:

- Privacy, so compute on-premise or decentralized
- Data is heavy, so compute on-premise
- Link in emerging decentralized AI compute

Objective function: Maximize supply of relevant *services*

=reward curating *services* + proving that it was delivered

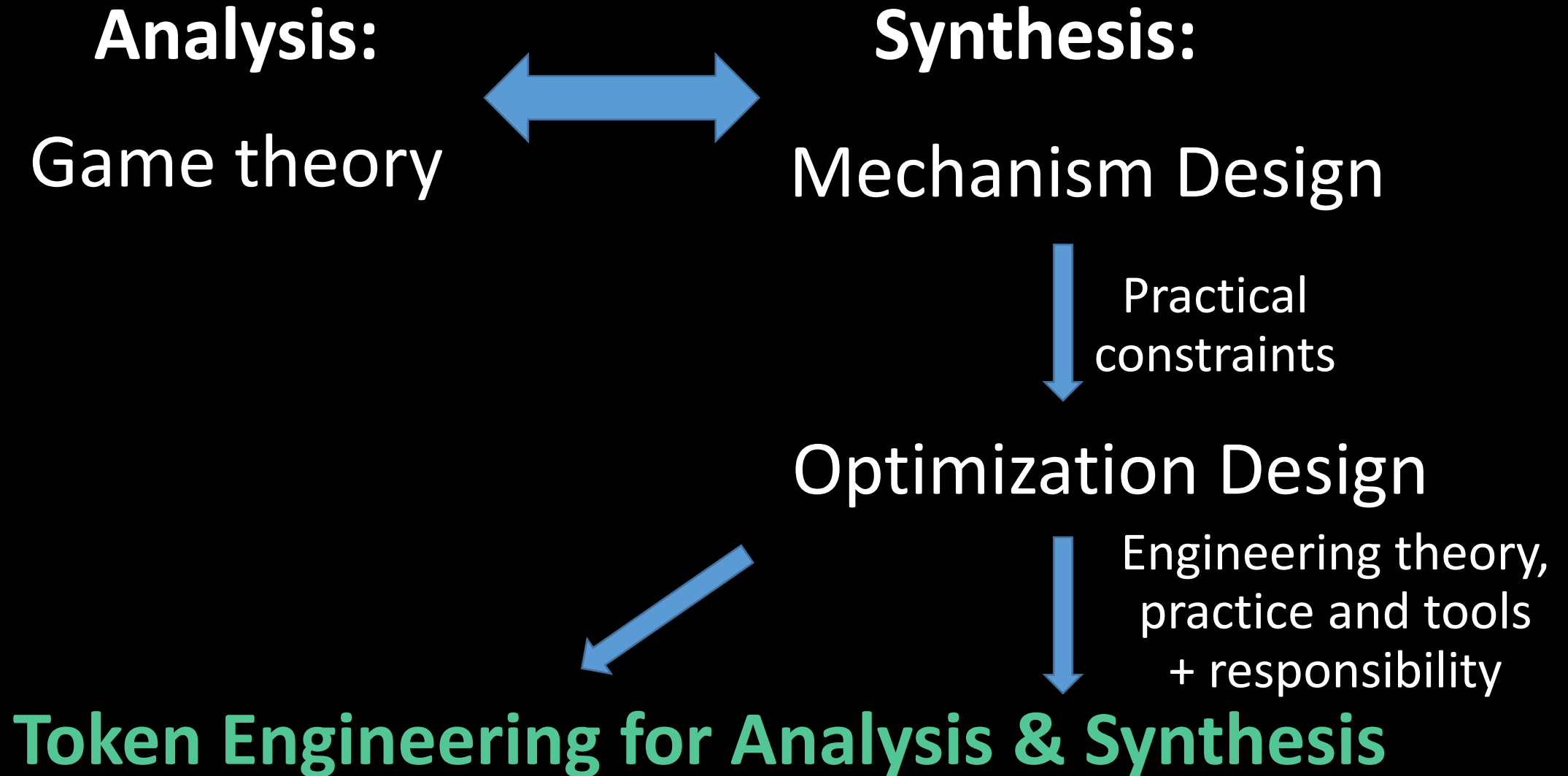
$$E(R_{ij}) \propto \underbrace{\log_{10}(S_{ij})}_{\text{predicted popularity of service}} * \underbrace{\log_{10}(D_j)}_{\text{proofed popularity of service}} * T * R_i$$

#Token Engineering



Design of Tokenized Ecosystems

From Mechanism Design to *Token Engineering*



#TokenEngineering has grown into a global movement

not secure | tokenengineering.net/community

site-name .wikidot.com Share on      

Expert tip #1: Wikidot modules let you make smarter web sites

[main](#) [discuss](#) [edit this page](#) [view source](#) [history](#) [o](#)

Community

Community: Events, Related Communities, more.

TE Local Meetup Groups

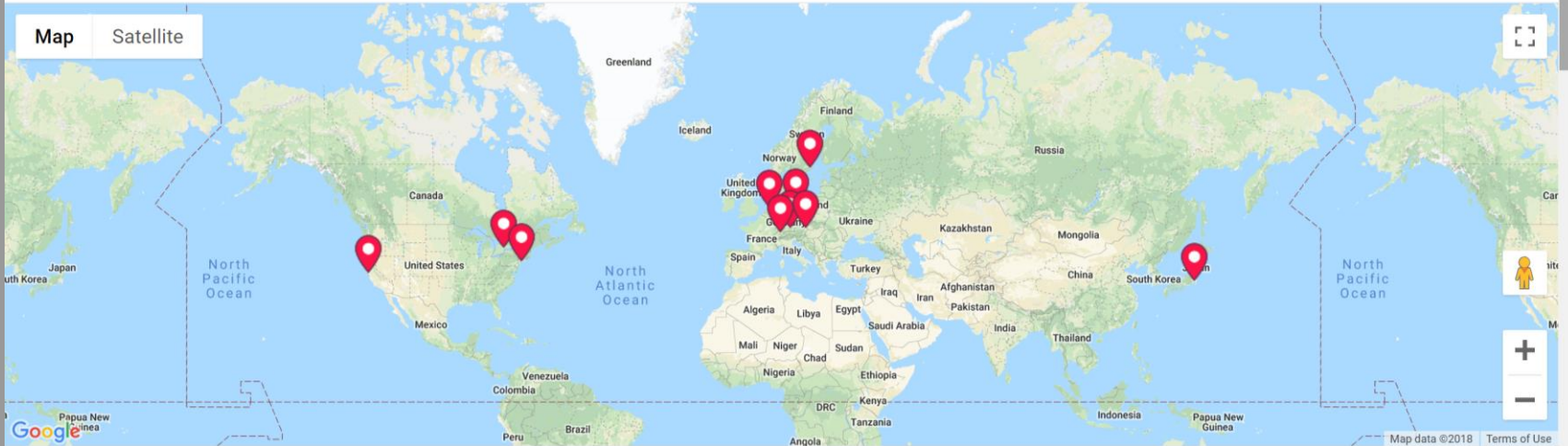
(The actual meetup.com pages will typically have the n

- [TE Berlin](#)
- [TE Budapest](#)
- [TE Munich](#)
- [TE Toronto](#)
- [TE London](#)
- [TE Zurich/Zug](#)
- [TE Tokyo](#)
- [TE NYC](#)
- [TE Amsterdam](#)
- [TE Stockholm](#)
- [TE San Francisco](#)
- [TE Vancouver](#)
- [TE Vienna](#)
- [TE St Petersburg](#)

Wanna start your own TE meetup? Please do! :) You c

 [Start a new group](#) [Explore](#) [Messages](#) [Notifications](#) 

[Map](#) [Satellite](#)



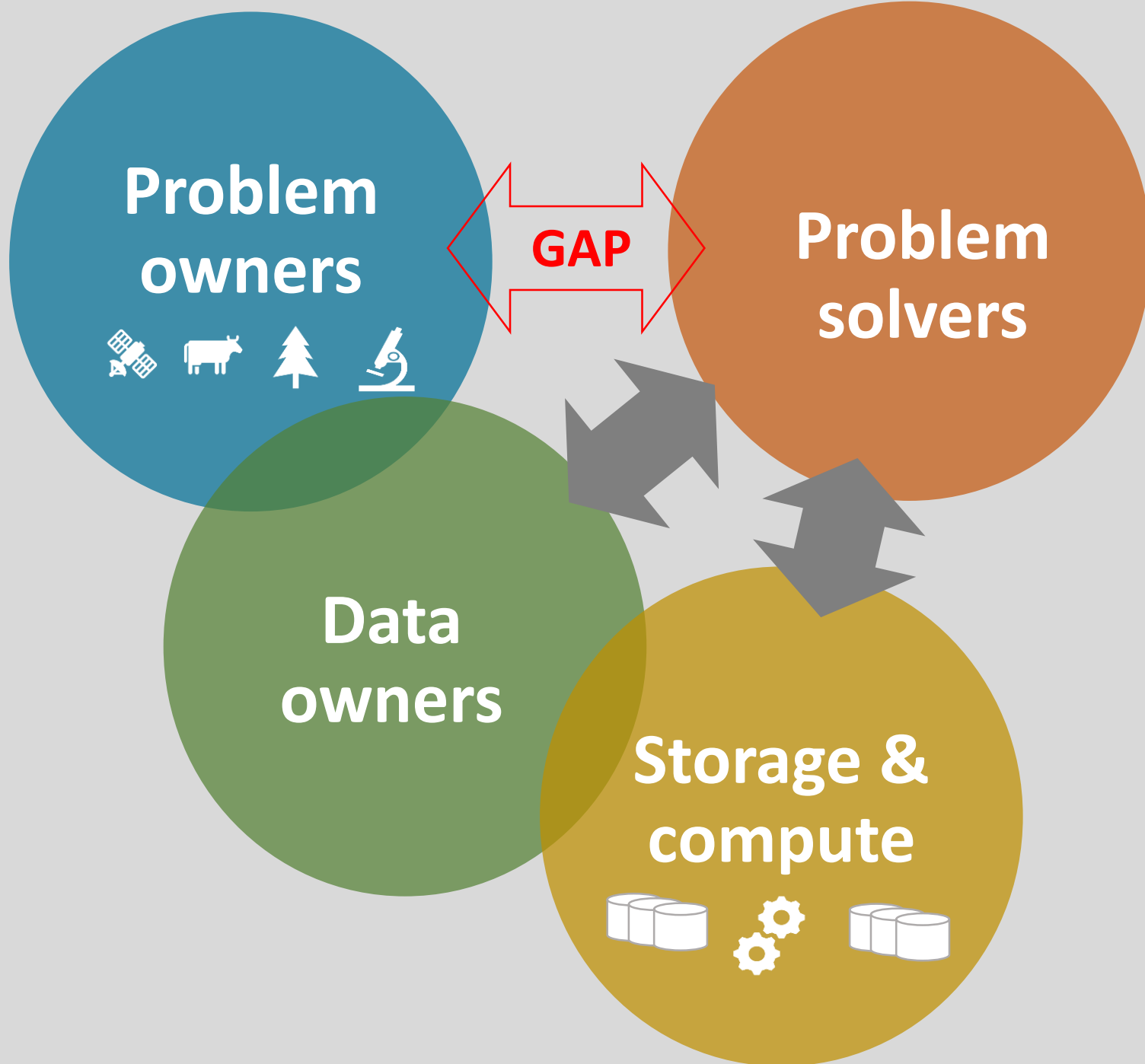
Google Map data ©2018 Terms of Use

 **Token Engineering Global**

Members	Groups	Countries
1,448	11	8

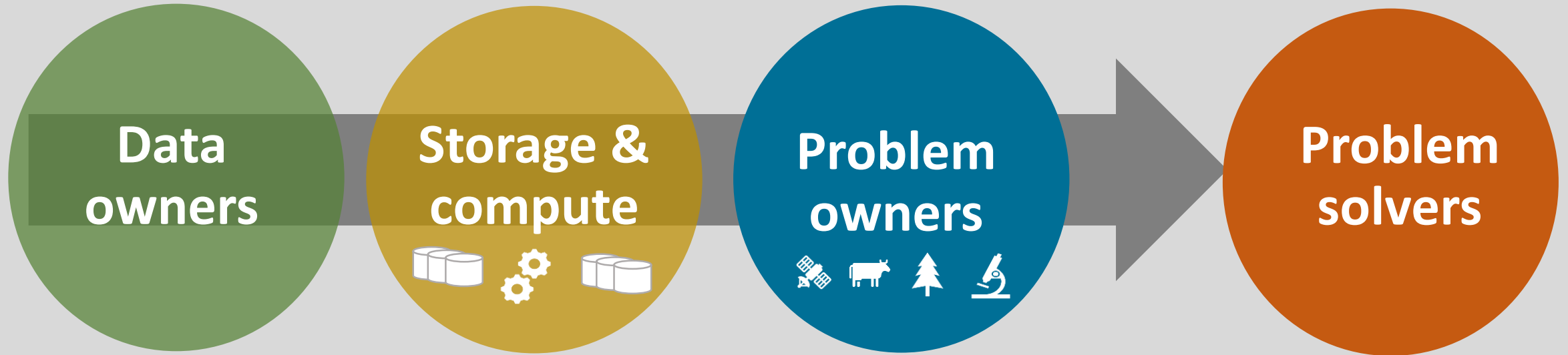
The background of the slide is a deep blue color, featuring several glowing jellyfish in various sizes and orientations. The jellyfish have a translucent, ethereal appearance with internal structures visible in shades of light blue and white. They are scattered across the frame, creating a sense of depth and movement.

Ocean, Towards a Decentralized Data & AI Stack

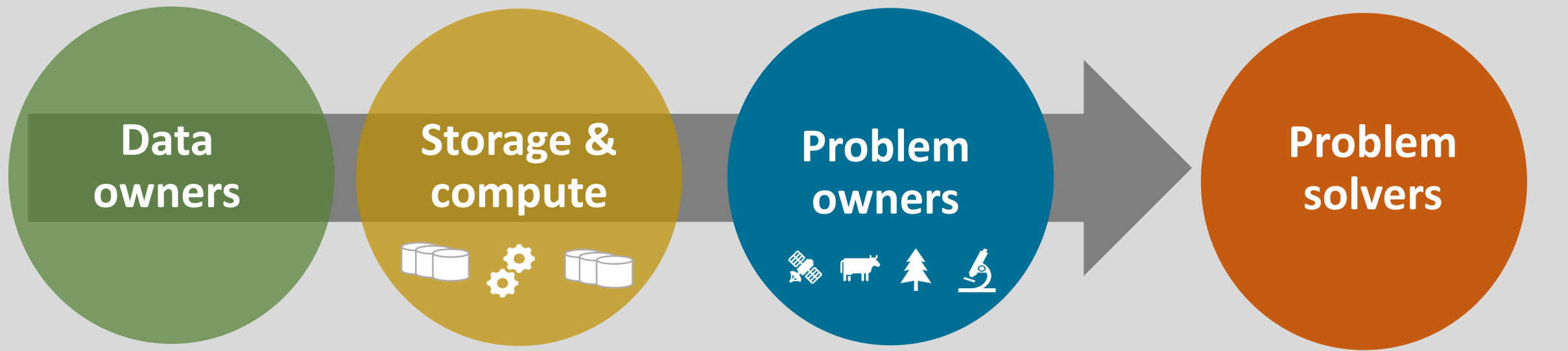




Connective tissue: public utility network



Blockchain public utility network for the AI Commons



Service provider frontends
For data commons,
Storage & compute markets

**AI Commons
Frontend**
Specify Problems

Data science tools

↓
**Data,
storage,
compute**

↑
**Network
rewards**

↓
**Jupyter
notebooks**

↑
**Notebooks,
data, storage,
compute**



Public utility network incentivizing data:

- Block rewards for data
- Privacy-preserving AI modeling

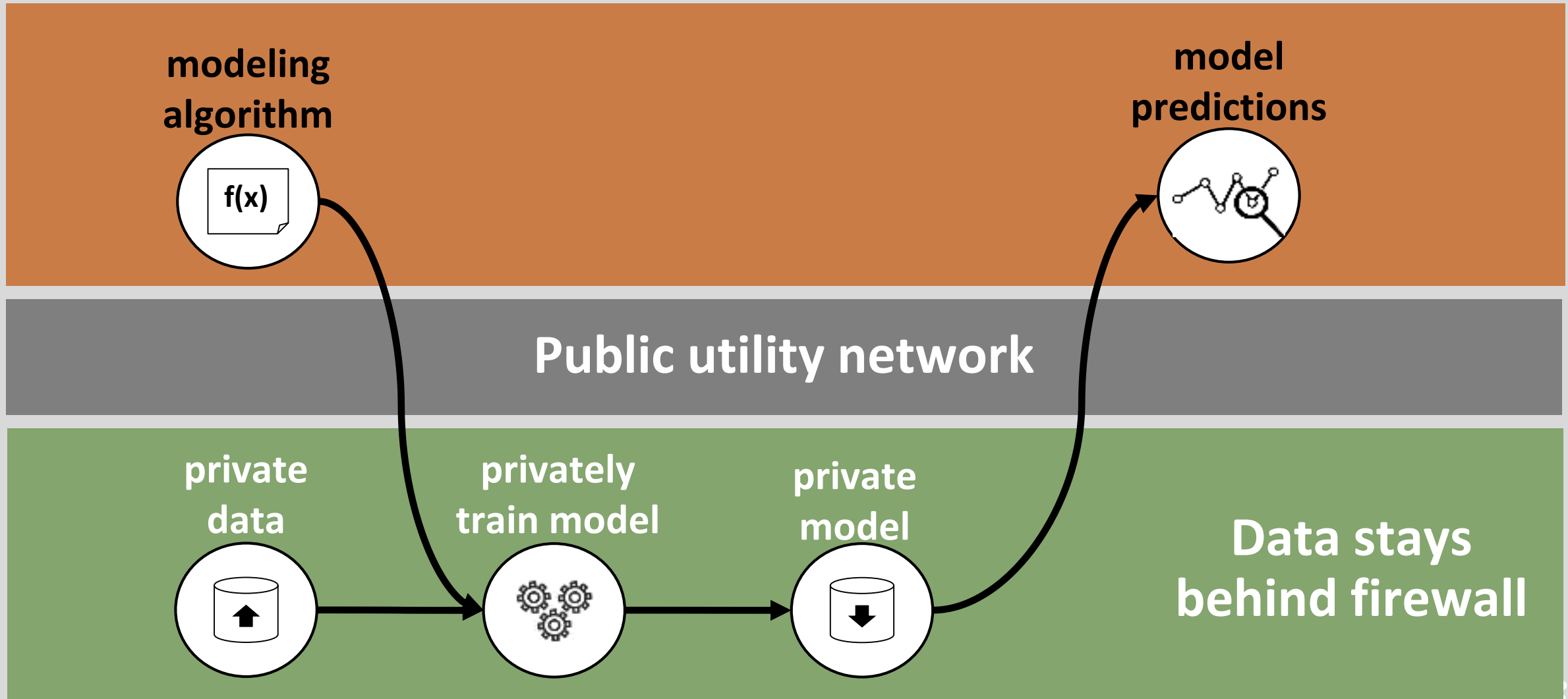
Have data	Have AI (Want data)	Have data	Have data	Have AI (Want data)
Have AI (Want data)	Have data	Have AI (Want data)	Have AI (Want data)	Have data
Data marketplace	DM	DM	DM	DM
DM	DM	DM	DM	DM
DM	DM	DM	DM	DM
DM	DM	DM	DM	DM

Catalyze the commons *and* profit-making, via a substrate for a thousand marketplaces to bloom.

Block rewards to incentivize data supply
Add to data commons, make \$



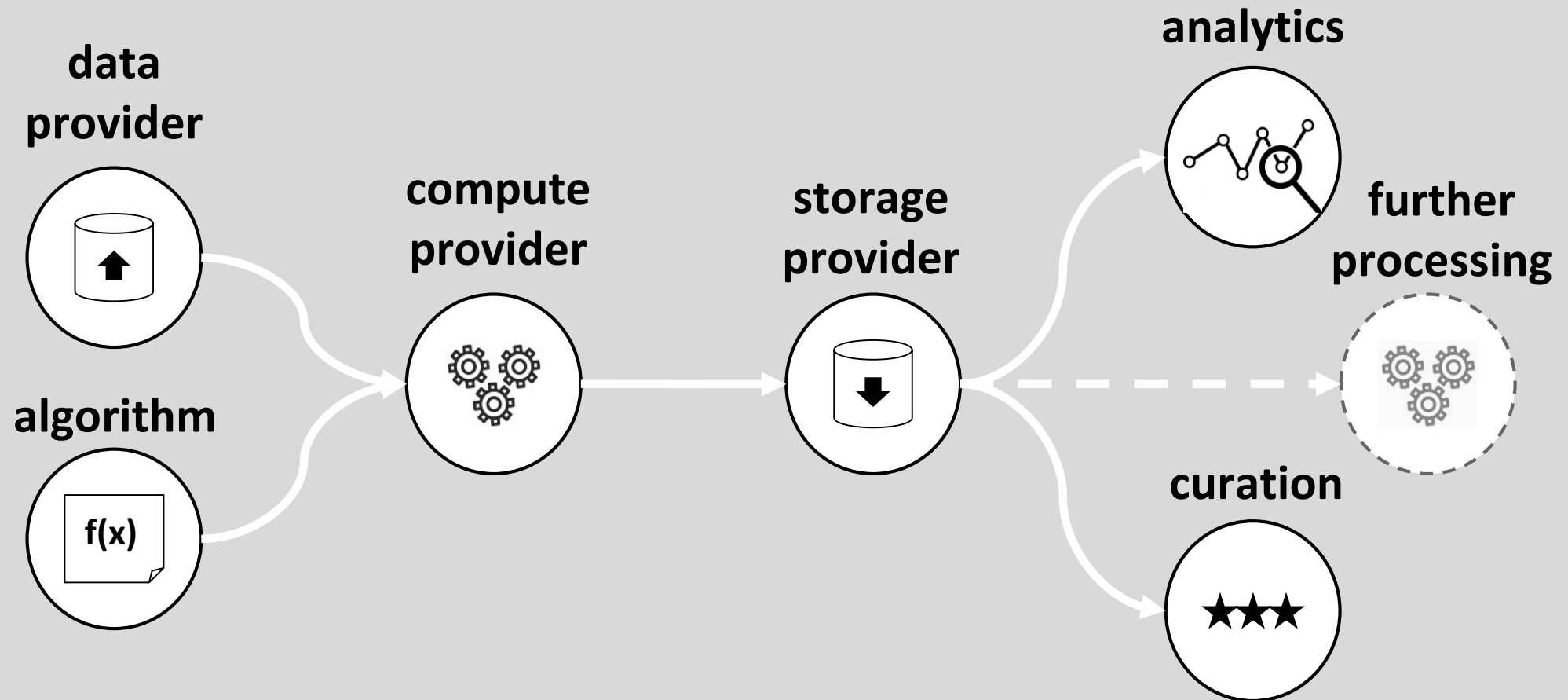
Preserve privacy: bring AI compute to the data



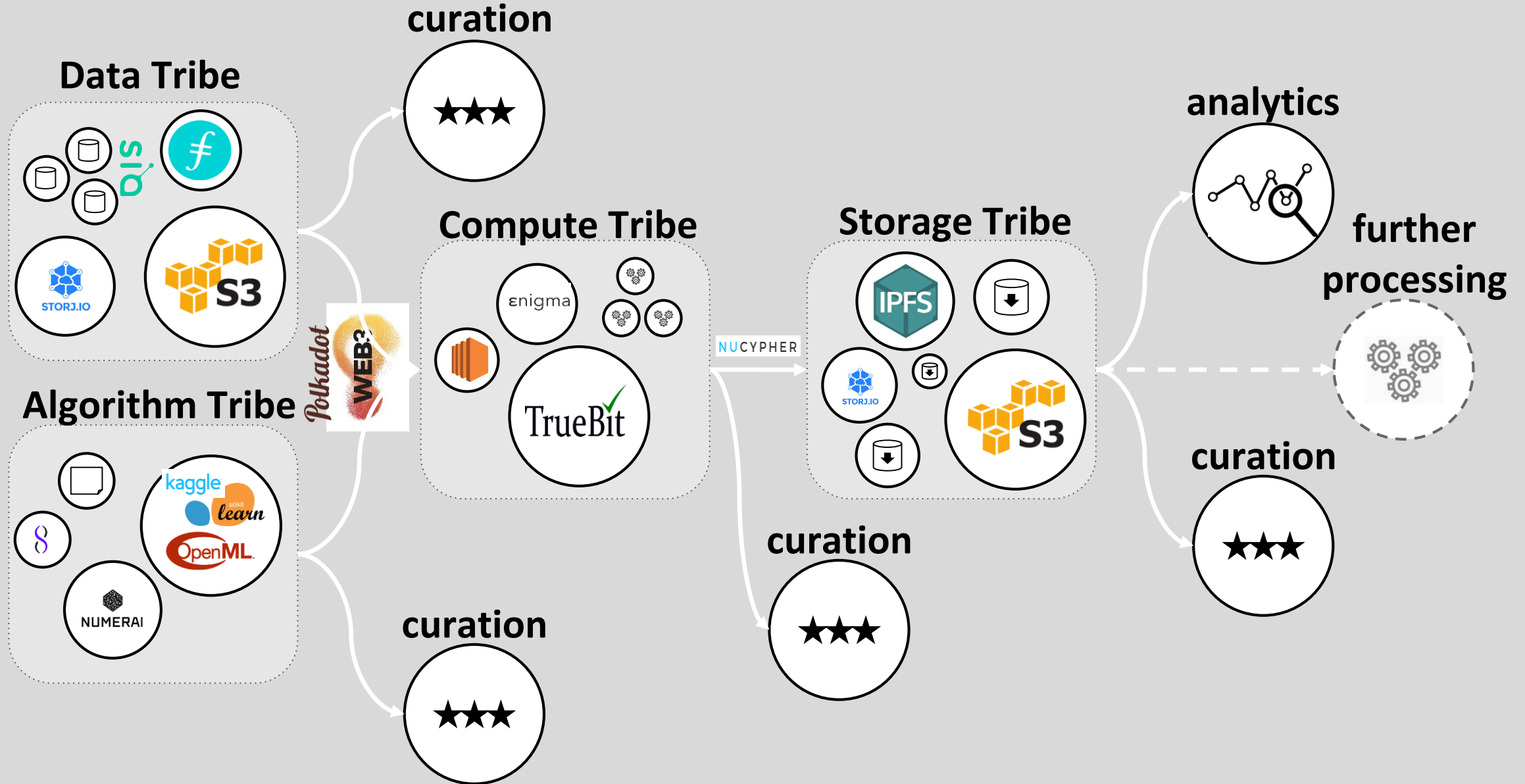
Framings for Ocean



“AI Compute Pipeline”

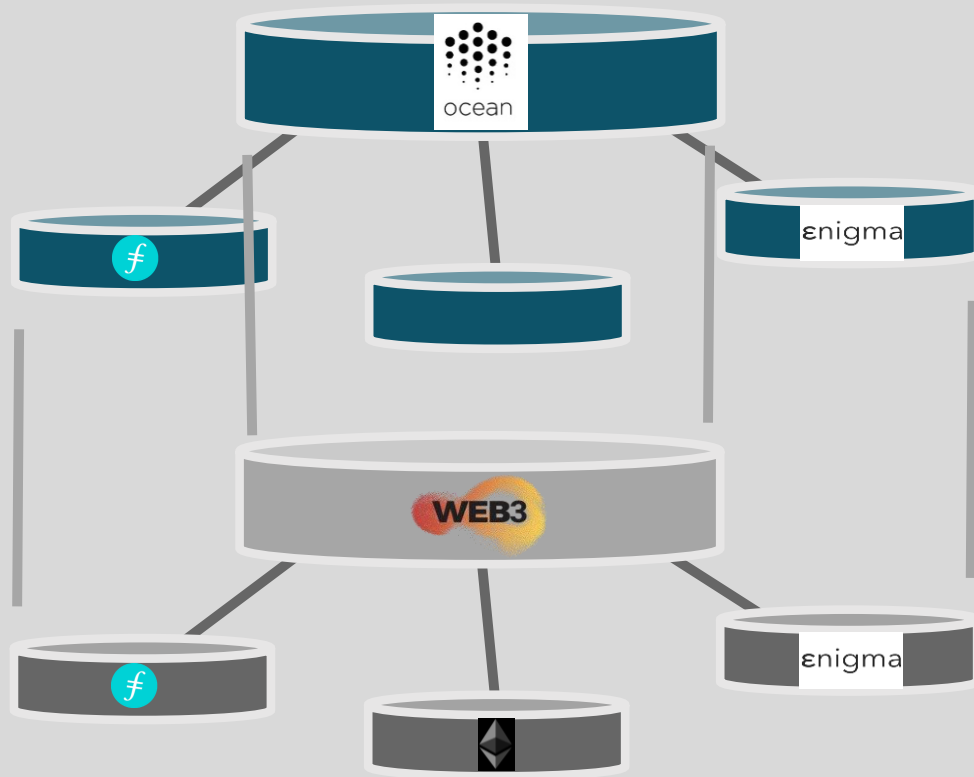


“Inter-Service Network”

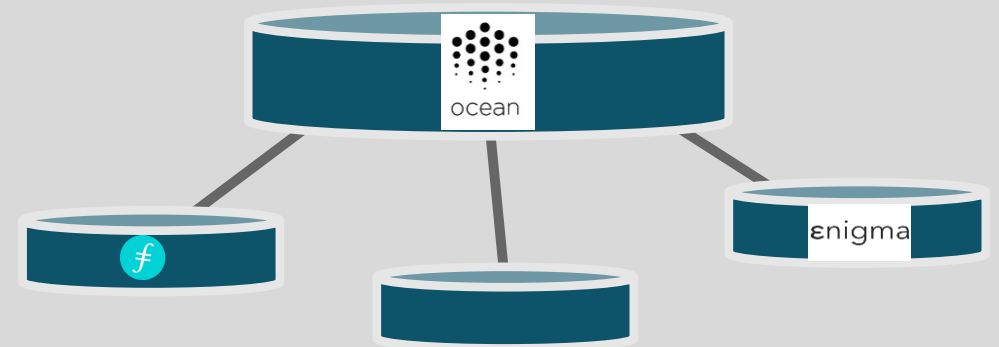


“Network of Networks, for AI Stack”

Crypto perspective:
Ocean is L2 on L1



AI perspective:
Ocean is L1



“SABRE for data”

SABRE is metadata/substrate for airline tickets. Ocean, for data itself.

Supply tickets:
Airlines

Consume tickets:
Consumer
travelers, biz ppl



Expedia

Kayak

3rd party
reseller

Easyjet
website

Travel
agent

Hip
munk

BA
website

Travel
agent

SABRE

Supply data:
Biz's, NGOs,
Governments

Consume data:
Data scientists,
Biz analysts



Data
market

DM

AC

DM

AI
common

DM

DM

AC

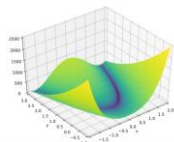
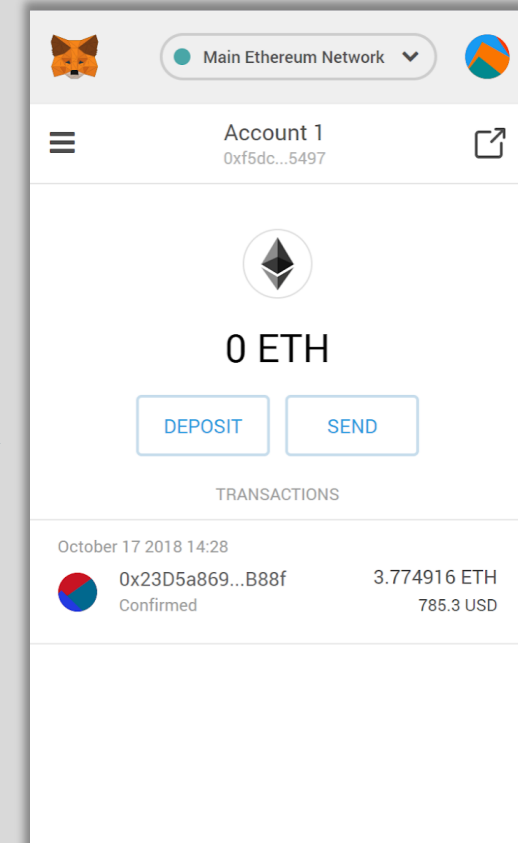
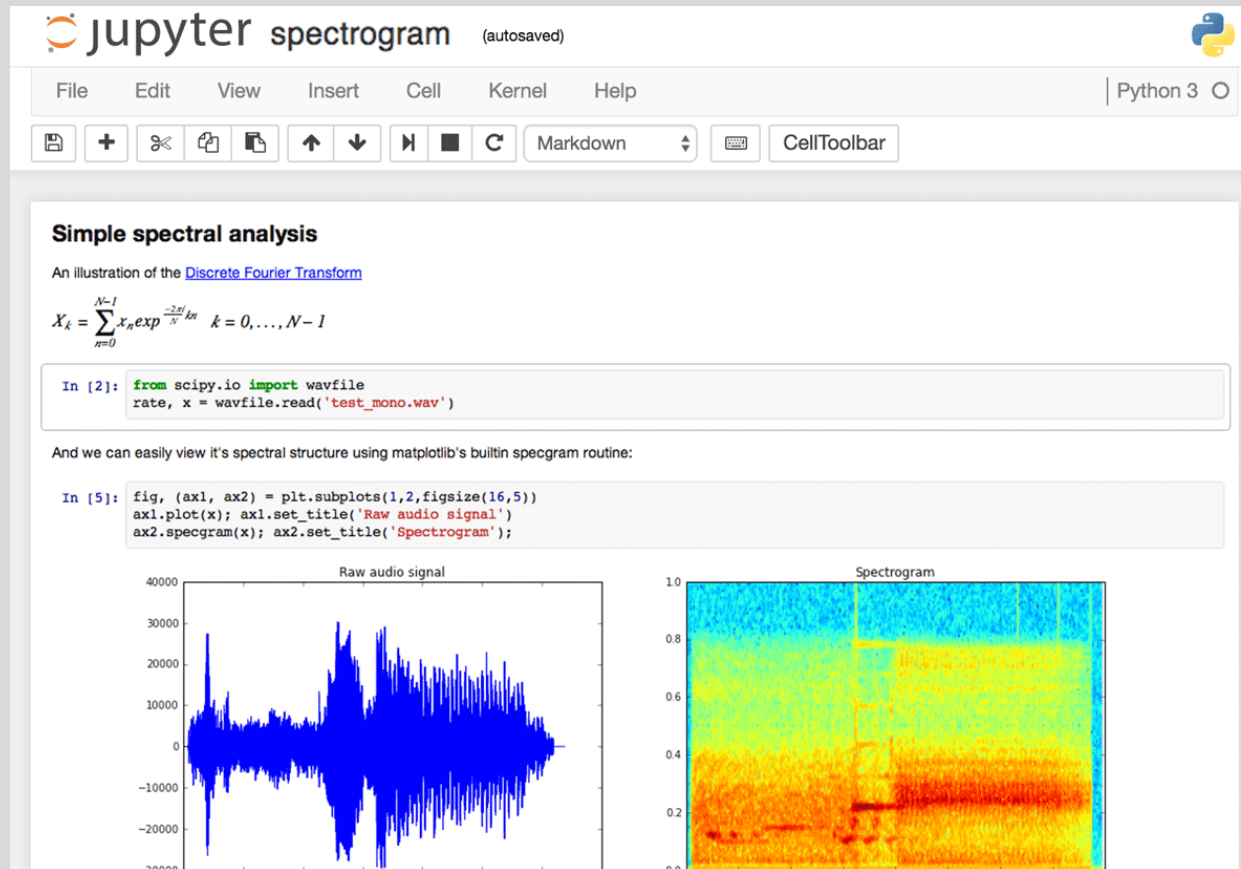


**Public utility
network**



**What does
decentralized AI look like?**

Keep Data Science Tooling, Add Crypto



Every dataset is an asset. Every Jupyter notebook is an asset (!)

Ocean ▶ KillTheHumans5 DatasetsViewLogin

Marketplace KillTheHumans

KillTheHumans is a marketplace for datasets and Jupyter notebooks. In order to eradicate piracy, all data must be input into the system using version 2.8.

1	Dataset 78XcYT9 Laser Target Practice Exercises	Publisher Name: LazerDrone Ltd. Date Published: 03-08-2019	LAZ
2	Dataset 87ertHG Facial Recognition Data	Publisher Name: VoyeurDrone Inc. Date Published: 20-06-2018	VOY
3	Dataset 000x567 Thermal Sensor Data	Publisher Name: Predator Movement Date Published: 31-09-2022	KTH
4	Dataset 96zERC Human Locations based on IP	Publisher Name: SkyNet Date Published: 09-11-2020	KLO
5	Dataset 254phyY Human Biological Weaknesses	Publisher Name: The Overmind Date Published: 28-06-2016	REX
6	Add your dataset here ... and become a publisher		

7845.12 OCN11 Datasets1 OCN = 0.12 ETHDiminator

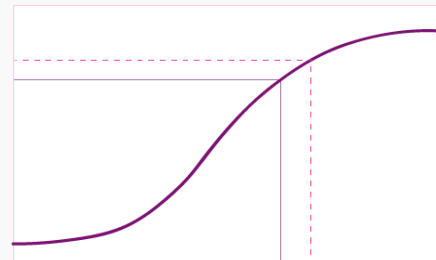
My Token Assets

NTT	102.10	KIO	55.00
RED	0.45	OSC	0.80
INT	45.38	IOX	1.00
ZAP	10.00	BAR	4523.12
NTT	100.00	URD	200.00
AAC	22.45		

My Dataset Portfolio

02	Dataset 0XhjijkA	▲ +1.8%	4.23k NTT	42.10 OCN
10	Dataset 0XhjijkA	▼ -0.5%	2.47k RED	01.10 OCN
04	Dataset qfij875	▲ +2.1%	5.89k INT	30.10 OCN
08	Dataset x00C475	▲ +5.2%	0.20k ZAP	42.10 OCN
02	Dataset 0XhjijkA	▼ -1.8%	4.23k NTT	40.10 OCN
07	Dataset 78XcYT9	▲ +4.5%	2.99k AAC	9.20 OCN
02	Dataset euyijC88	▲ +1.8%	4.23k KIO	38.26 OCN
10	Dataset 78eyujP	▼ -4.6%	0.45k OSC	42.10 OCN
19	Dataset 0XhjijkA	▼ -2.3%	4.23k IOX	22.47 OCN
22	Dataset AA854cv	▲ +0.4%	6.01k BAR	42.10 OCN
03	Dataset 0XhjijkA	▲ +1.8%	8.23k URD	10.24 OCN

Dataset Selection and Token Supply vs Cost

Dataset 78XcYT9	Marketplace #Name	Ranking 7
 ▲ +4.5% Current 9.20 OCN 2.99k AAC Future 9.25 OCN 3.04k AAC		
◀ +5.80 ▶ Invest (+) or Withdraw (-) Confirm Transaction		

New powers for data scientists

- **Way more data.** Data commons. Enterprise data without data escapes via on-premise compute.
- **Provenance** in data & AI training. Goodbye data honeypots.
- **More \$.** For generating data. For cleaning, labeling, feature engineering data. For algs. For curation.

https://docs.oceanprotocol.com

Ocean Protocol Documentation - x

https://docs.oceanprotocol.com



ocean

Ocean Protocol Documentation

Learn about the components of the Ocean Protocol software stack, and how to run or use the components relevant to you.

Core Concepts

Understand the fundamentals of Ocean Protocol.

[Learn More →](#)

Setup Guides

Setting up the Ocean Protocol components.

[Learn More →](#)

Tutorials

Browse tutorials for most common setup and development use-cases.

[Learn More →](#)





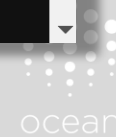
Manta Ray

Data Science powered by Ocean Protocol

The **Manta Ray** notebooks provide a guided tour of Ocean Protocol in an interactive Jupyter Notebook environment. Start using Ocean Protocol with your own pre-configured and loaded cloud instance after login with your GitHub account.

This project is in alpha! Feel free to ask questions and post bug reports in [our Gitter channel](#). Notebooks are for tutorial and demonstration purpose only. Notebook instances may be periodically offline, and storage volumes will be purged.

JupyterLab Instance



QUICKSTART

Run & Try Everything

MARKETPLACES

Set Up a Marketplace

PUBLISHERS

Publish Data or Services

KEEPERS

Run a Keeper

```
git clone https://github.com/oceanprotocol/barge.git
cd barge/

./start_ocean.sh --latest
```

Seeing the dolphin means it's working:

```
→ ./start_ocean.sh --latest
```

```

                                     Welcome to

                                     _.-'-.
                                   .'.'.-'.'-'.
                                 /.'-.'--'-'-'-'-' \
                               ;/         \_-'-'-'-' \
                              /_-;        /_-'-'-'-' \

      _.-._)'\_.-._)'\_.-._)'\_.-._)'\_.-._
     /___\ /___\ /___\ /___\ /___\ /___\ /___\ /___\ /___\
    |  ()  |  - )  |  ' \  |  ' \  |  ' \  |  ' \  |  ' \  |
    \___/ \___/ \___/ \___/ \___/ \___/ \___/ \___/ \___/

                                     ---

Get info: oceanprotocol.com
Get code: github.com/oceanprotocol
Get help: gitter.im/oceanprotocol/Lobby
```



Conclusion

Conclusion:

On Decentralizing AI Data & Services

- Goal: spread benefits of AI, by equalizing opportunities to access data
- Token design $\approx$ EA design.
 - So, approach token design as EA design!
 - Helped a *lot* for designing Ocean. #TokenEngineering.
- Decentralized AI gives data scientists new powers
 - *Way* more data. Privacy-preserving compute.