ManaTEE: an Open-Source Private Data Analytics Framework with Confidential Computing

Dayeol Lee

dayeol.lee@tiktok.com

Contents

- Private Data Analytics
  - Why is it Hard?
  - What do we Need?
- Existing Approaches
- ManaTEE Project
- Tutorial Demo
- Project Current & Future

Private Data Analytics

Why is Private Data So Important?

Value Extraction

Public Interest

Why is Private Data So Important?

Value Extraction

Public Interest

Sharing Private Data for Public Interest

- **Public Health**: medical data, personal health data
- **Public Safety**: PII (e.g., address, phone number) associated with crimes or illegal activities
- **Education**: academic performance, attendance, and engagement
- **Civic Engagement**: personal beliefs, social activities

… and many more

Often Requires Cross-Organizational Data

Example: Understanding Illicit Drug Promotion by Using Cross-Platform Data
(Zha et al., CCS’24)

Understanding Cross-Platform Referral Traffic for Illicit Drug Promotion

Mingming Zha
mzha@iu.edu
Indiana University Bloomington
Bloomington, Indiana, USA

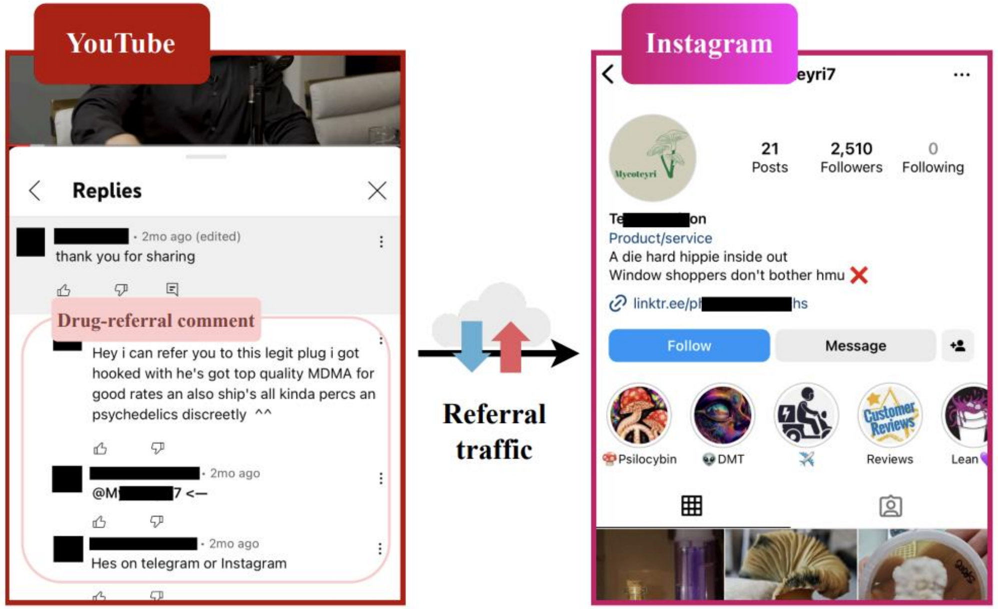
Zilong Lin
zillin@indiana.edu
Indiana University Bloomington
Bloomington, Indiana, USA

Siyuan Tang
tangsi@iu.edu
Indiana University Bloomington
Bloomington, Indiana, USA

Xiaojing Liao
xliao@indiana.edu
Indiana University Bloomington
Bloomington, Indiana, USA

Yuhong Nan
nanyh@mail.sysu.edu.cn
Sun Yat-sen University
Zhuhai, Guangdong, China

XiaoFeng Wang
xw7@indiana.edu
Indiana University Bloomington
Bloomington, Indiana, USA



[redacted] · 2mo ago (edited)
thank you for sharing

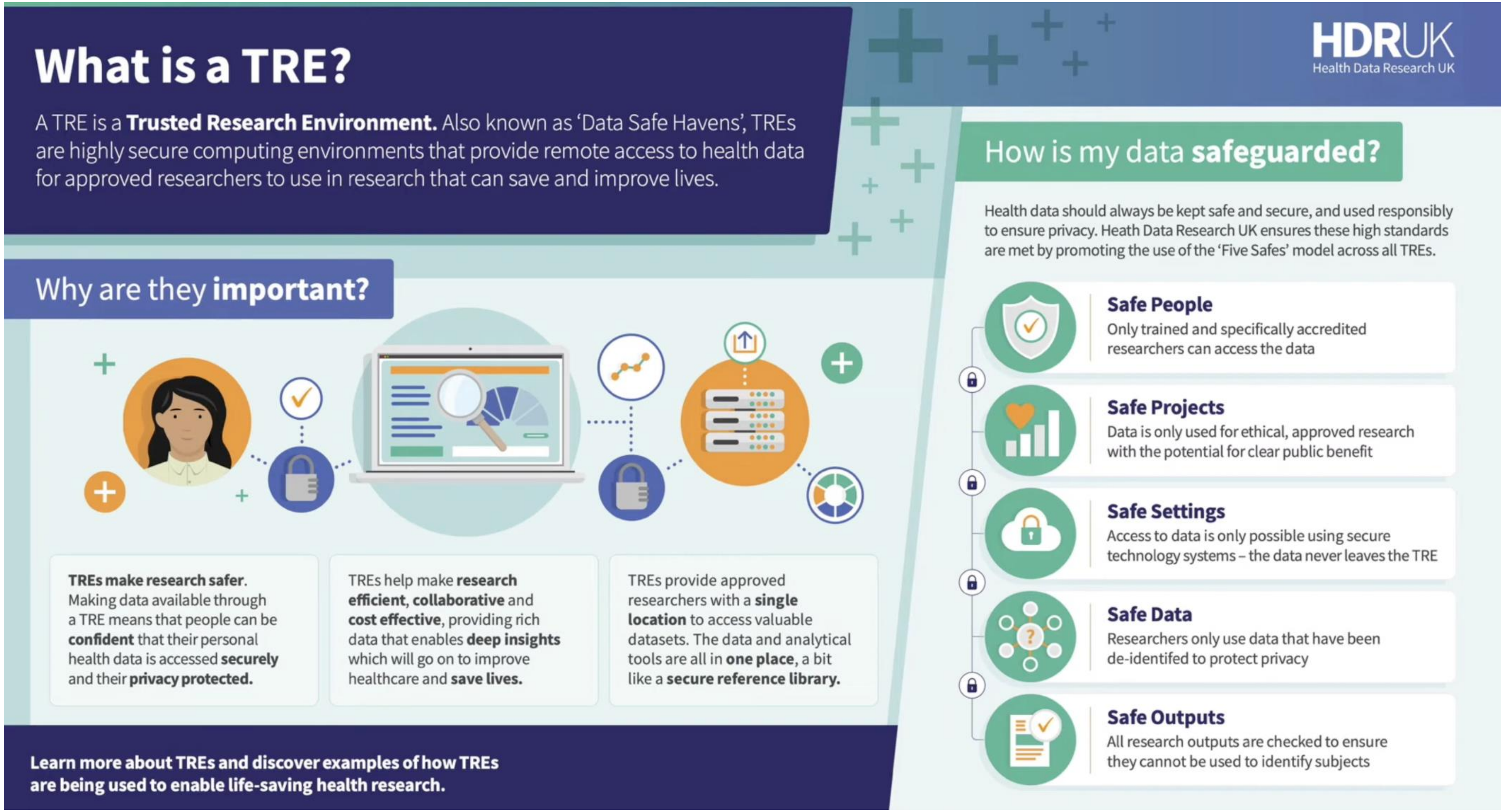
[redacted] Hey i can refer you to this legit plug i got hooked with he's got top quality MDMA for good rates an also ship's all kinda percs an psychedelics discreetly ^^

[redacted] · 2mo ago
@M[redacted]7 <—

[redacted] · 2mo ago
Hes on telegram or Instagram

eyri7
21 Posts 2,510 Followers 0 Following
Te[redacted]on
Product/service
A die hard hippie inside out
Window shoppers don't bother hmu ❌
linktr.ee/pl[redacted]hs

HDR UK: Trusted Research Environment (TRE)



HDRUK
Health Data Research UK

What is a TRE?

A TRE is a **Trusted Research Environment.** Also known as 'Data Safe Havens', TREs are highly secure computing environments that provide remote access to health data for approved researchers to use in research that can save and improve lives.

Why are they **important?**

TREs make research safer. Making data available through a TRE means that people can be **confident** that their personal health data is accessed **securely** and their **privacy protected.**

TREs help make **research efficient**, **collaborative** and **cost effective**, providing rich data that enables **deep insights** which will go on to improve healthcare and **save lives.**

TREs provide approved researchers with a **single location** to access valuable datasets. The data and analytical tools are all in **one place**, a bit like a **secure reference library.**

Learn more about TREs and discover examples of how TREs are being used to enable life-saving health research.

How is my data **safeguarded?**

Health data should always be kept safe and secure, and used responsibly to ensure privacy. Heath Data Research UK ensures these high standards are met by promoting the use of the 'Five Safes' model across all TREs.

Safe People
Only trained and specifically accredited researchers can access the data

Safe Projects
Data is only used for ethical, approved research with the potential for clear public benefit

Safe Settings
Access to data is only possible using secure technology systems – the data never leaves the TRE

Safe Data
Researchers only use data that have been de-identifed to protect privacy

Safe Outputs
All research outputs are checked to ensure they cannot be used to identify subjects

https://www.hdruk.ac.uk/access-to-health-data/trusted-research-environments/

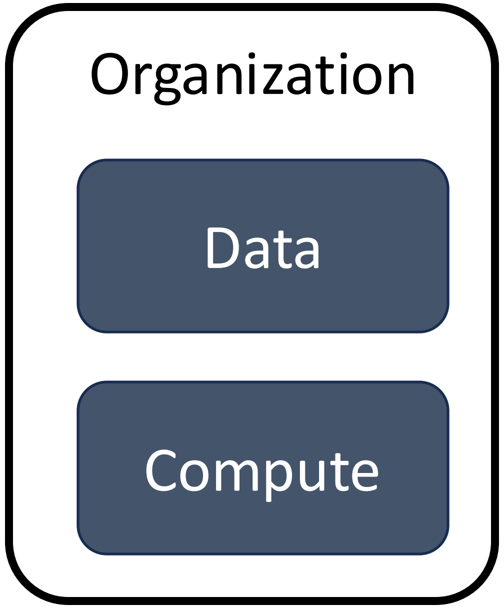
Why is it Hard?

Challenge 1: Data Privacy Risks

- Trust
  - Conflict of interests
  - Risk of abusing data or data fabrication
  - Different trust domain (e.g., company, country, …)
- Compliance
  - Privacy policies such as data retention or purpose limitation
  - Providing raw data might be legally prohibited
  - Changing the geolocation of data can be legally restricted

Challenge 2: Accountability and Transparency

In the Old Days…

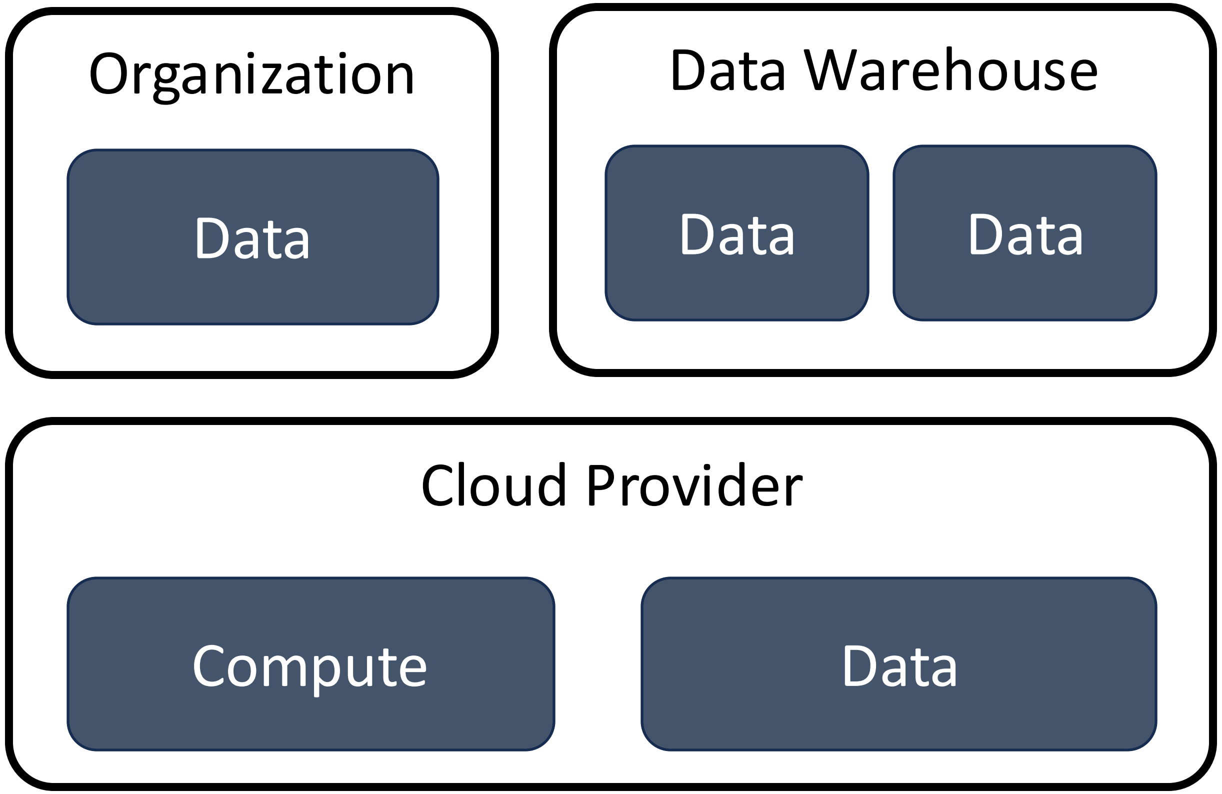


Organization

Data

Compute

Now it looks like



Organization

Data

Data Warehouse

Data

Data

Cloud Provider

Compute

Data

What do we Need?

We Need A “Standard” Way That Provides…

- Strong **Privacy Protection** Mechanisms
  - Privacy Enhancing Technologies (PETs)
  - Protecting privacy while maximizing the utility of data
- **Technical Enforcement** of Policies
  - Terms and conditions and honor codes are not enough
  - Proactive measures, instead of reactive
- **Accountability** and **Transparency**
  - Auditability of the full system if necessary
  - Verifiability on the integrity of the results
- **Usability**
  - Provide accurate results
  - Must be easy to deploy, easy to use, and easy to customize

Existing Approaches

Existing Industry Solutions for Security & Privacy

SQL Policy-based Data Clean Room

Rely on SQL/data platforms provided by 3rd party, who is free of conflict of interest

Differential Privacy

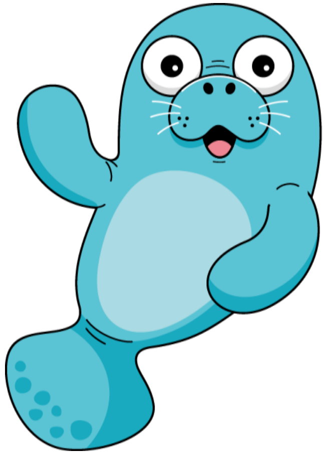
Preprocess data or add noise to the aggregated SQL results to limit information leakage

Trusted Execution Environment

Use remote attestation to co-verify the code before releasing data; contain data in an isolated environment during execution

Technical Difficulties of Existing Solutions

	PET	Technical Enforce	Transparency	Usability	Accuracy
SQL Policy-based Data Clean Room	No	No	No	Good	Yes
Differential Privacy	Yes	Yes	No	Good	No
Trusted Execution Environment	Yes	Yes	Yes	Could Be Better	Yes

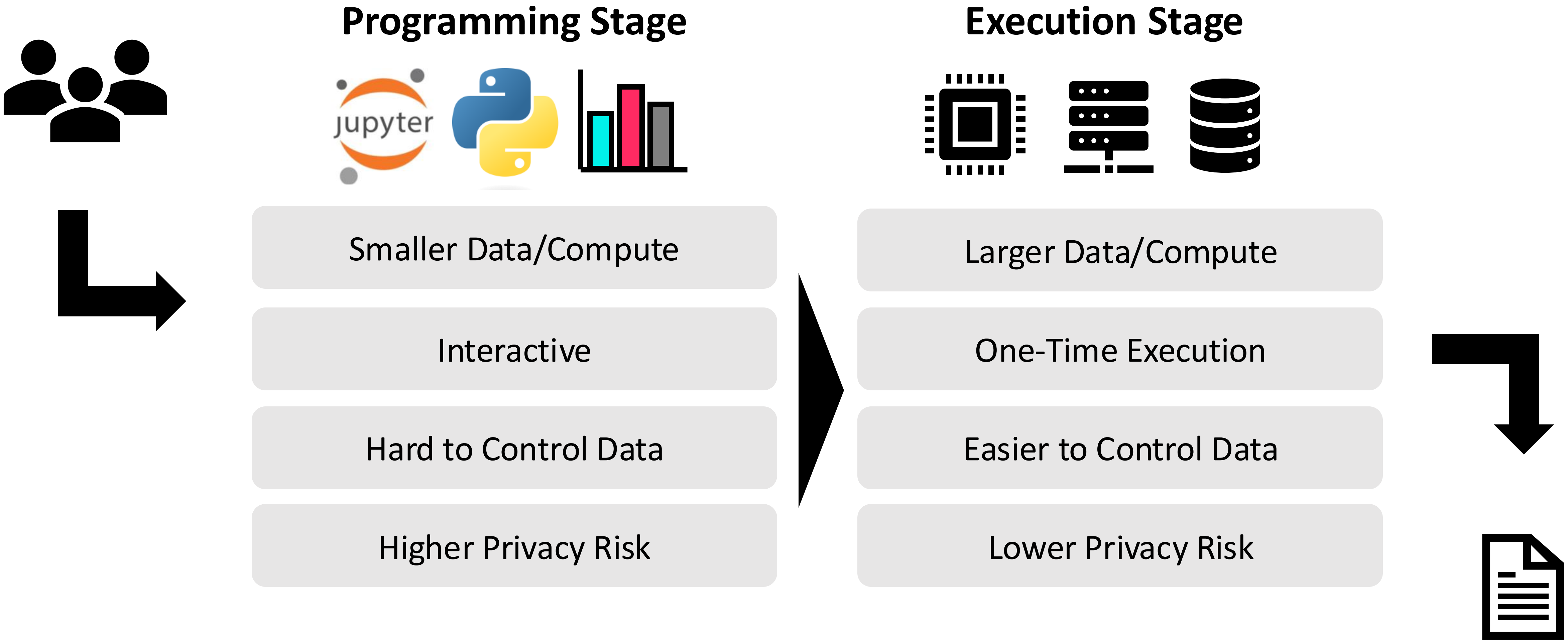


ManaTEE

Our Goals

- **Technical Enforcement via PET**: enforce privacy policies such as purpose limitation and data retention via PETs
- **Usability**: provide an interactive tool to utilize the data
- **Accuracy**: provide accurate results on real data, as well as an evidence of execution
- **Transparency and Accountability**: make it auditable and verifiable
- **Deployment**: make it easy to deploy to the cloud

Different Need at Each Stage



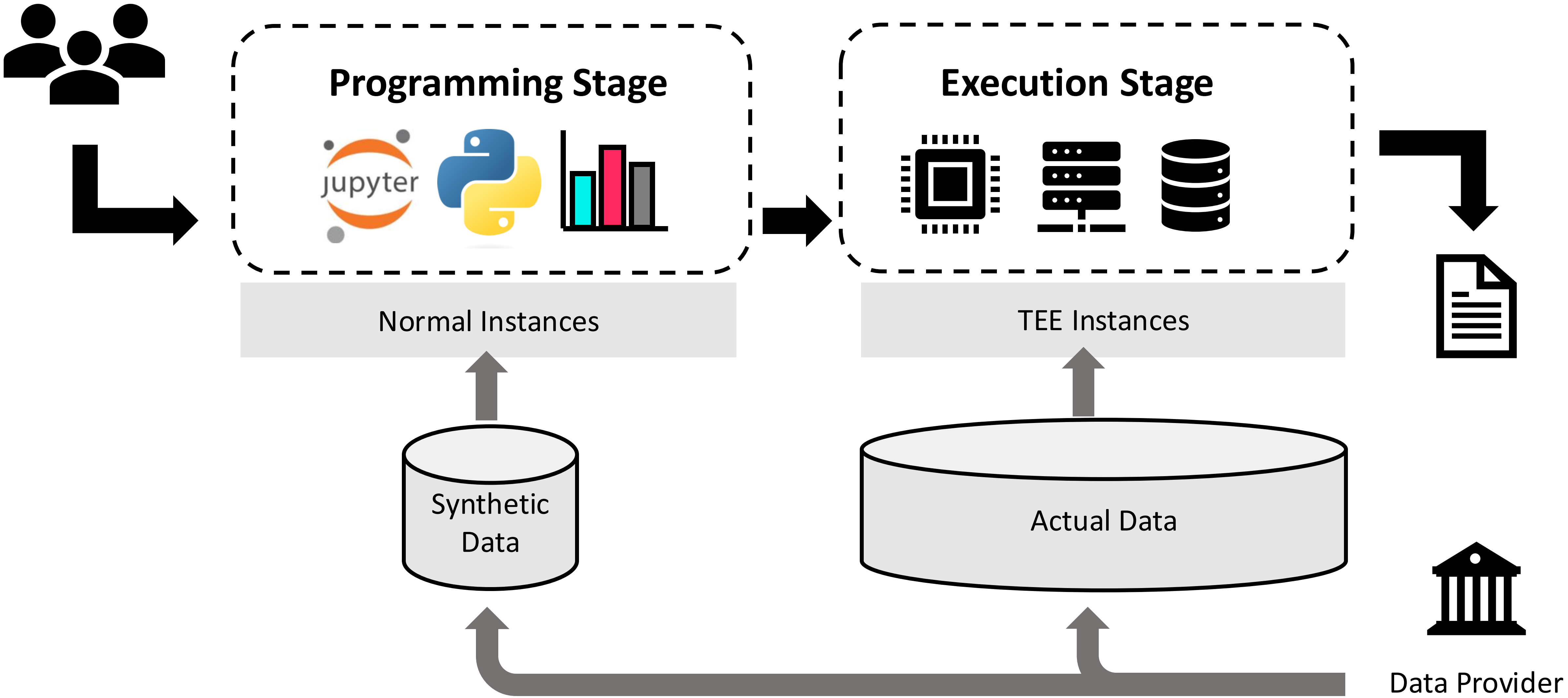
Programming Stage

- Smaller Data/Compute
- Interactive
- Hard to Control Data
- Higher Privacy Risk

Execution Stage

- Larger Data/Compute
- One-Time Execution
- Easier to Control Data
- Lower Privacy Risk

Our Approach: Two-Stage Data Clean Room



Programming Stage

Execution Stage

Normal Instances

TEE Instances

Synthetic Data

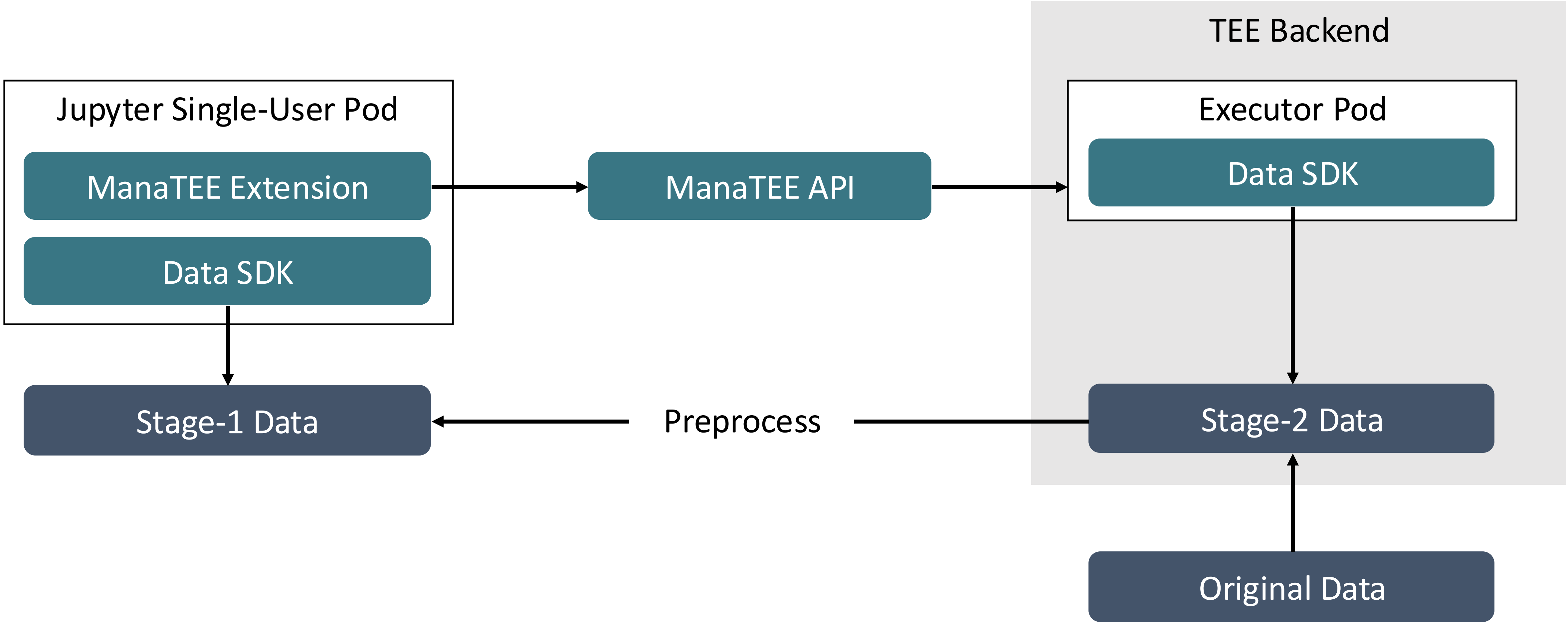
Actual Data

Data Provider

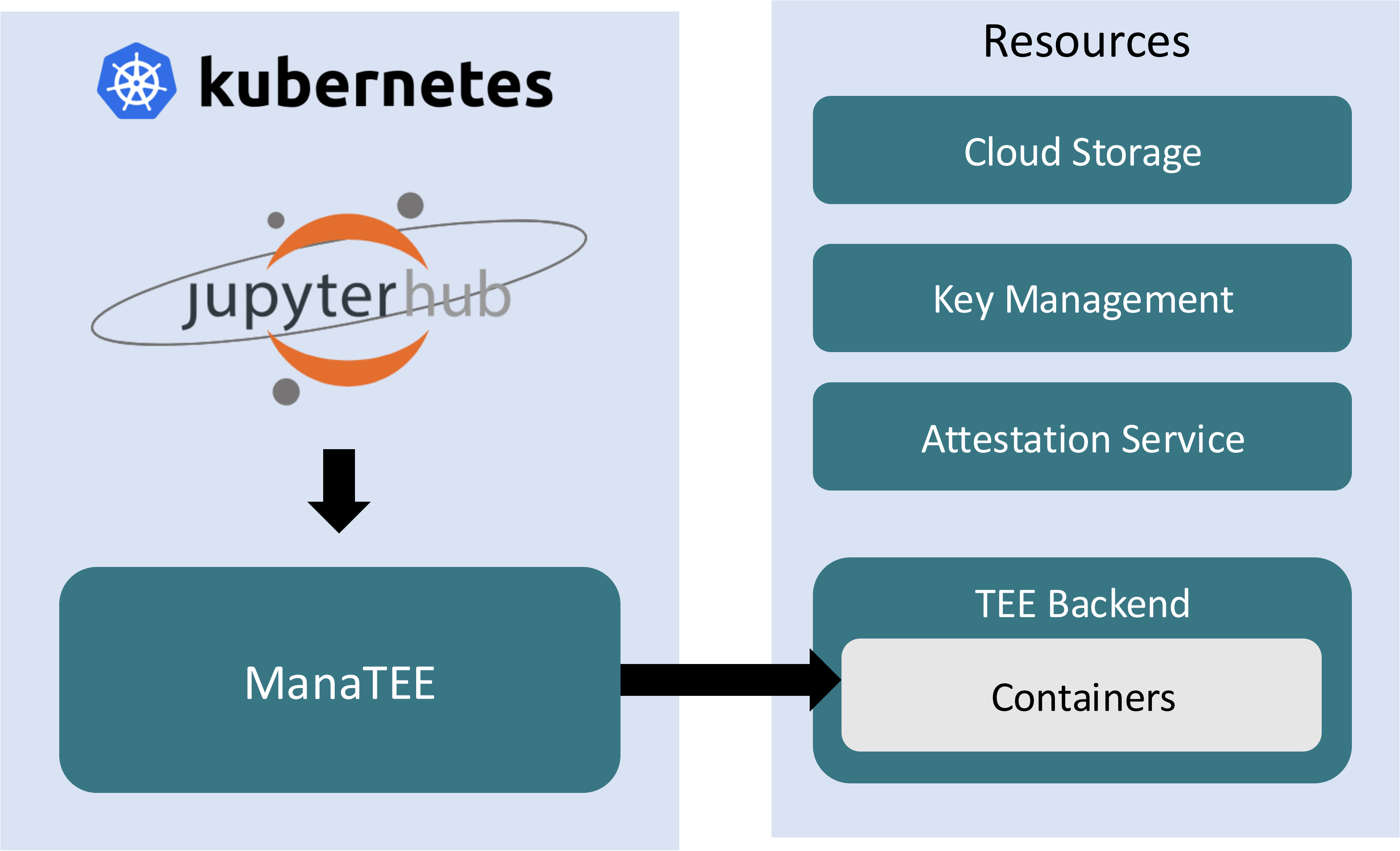
Benefits of the Two-Stage Execution Model

- Separate **Data Policy** and **Code Policy**
  - Flexible data policy on programming stage – LDP perturbation, sampled data, or DP synthetic data
  - Code policy enforced only at the execution stage
- **Accurate** Results in Execution Stage
  - Full data access is securely enabled via confidential computing
- Why **Confidential Computing**?
  - Provides transition of trust, making it work with various trust model (Cross-organizational data providers)
  - Integrity of the execution
  - Proof of execution (attestation report)

ManaTEE Data and Code Pipeline

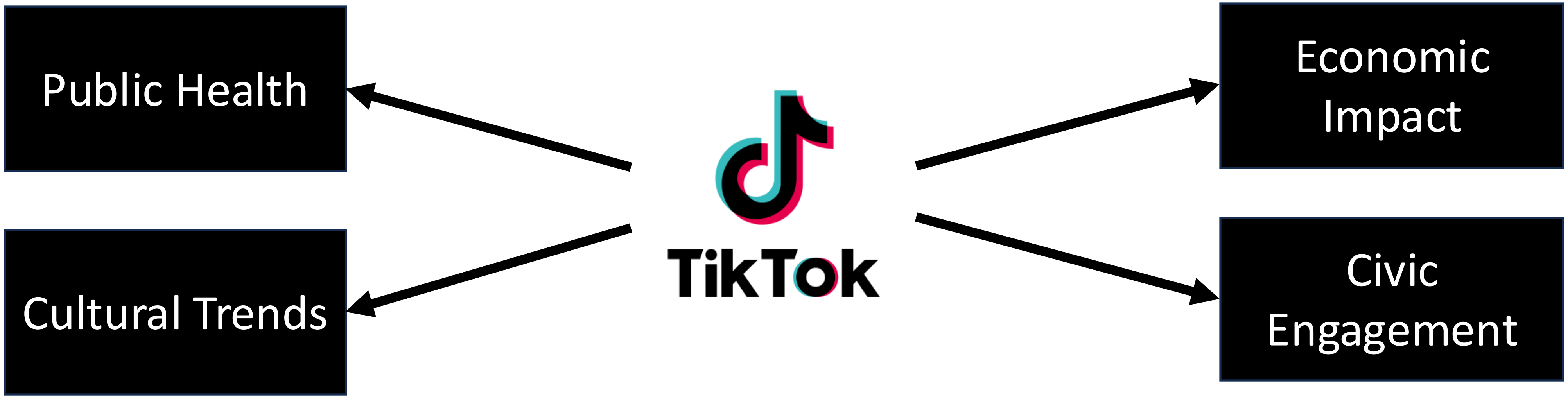


Easy Cloud Deployment via Terraform



Use Cases

Providing Transparency to Researchers



Public Health

Cultural Trends

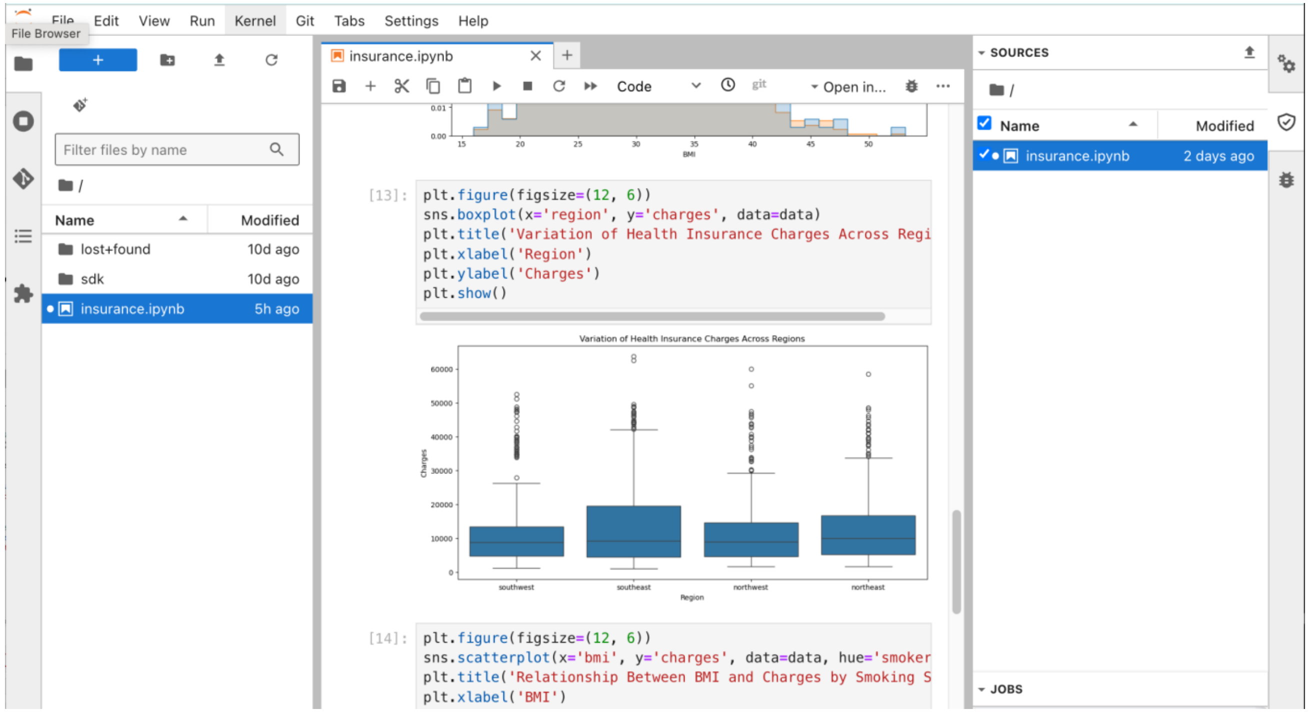
TikTok

Economic Impact

Civic Engagement

TikTok Research Tools

- Virtual Compute Environment (VCE) https://developers.tiktok.com/doc/vce-getting-started



```
[13]: plt.figure(figsize=(12, 6))
      sns.boxplot(x='region', y='charges', data=data)
      plt.title('Variation of Health Insurance Charges Across Regi
      plt.xlabel('Region')
      plt.ylabel('Charges')
      plt.show()
```

```
[14]: plt.figure(figsize=(12, 6))
      sns.scatterplot(x='bmi', y='charges', data=data, hue='smoker
      plt.title('Relationship Between BMI and Charges by Smoking S
      plt.xlabel('BMI')
```

Another Potential Use Cases

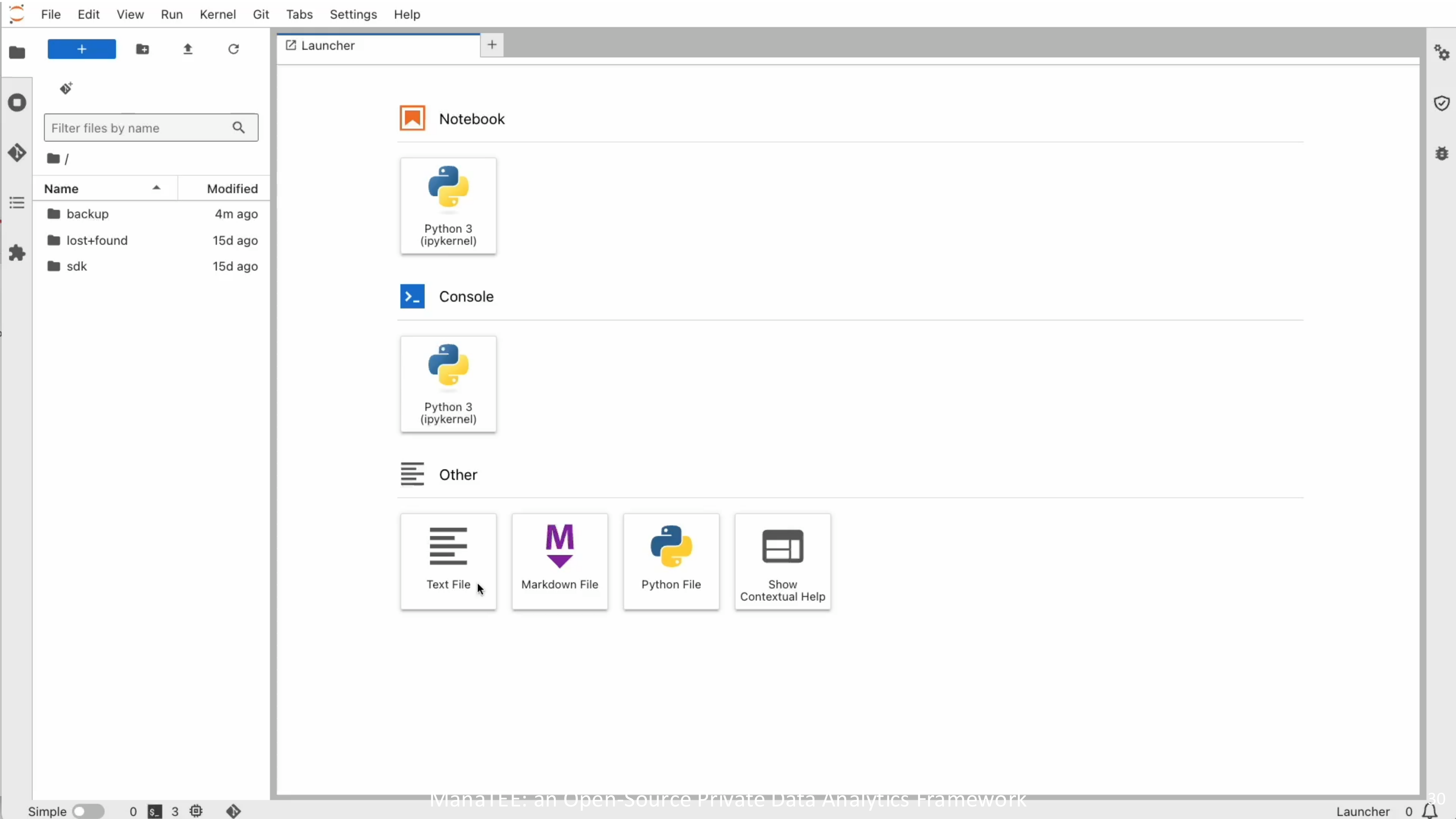
- Ads & Marketing
  - Lookalike segment analysis
  - Measurement and conversion tracking
- Machine Learning
  - Inferencing & training with private dataset
  - Inferencing & fine-tuning private model
  - AI model evaluation (e.g., fairness) on private models

Tutorial Demo

https://manatee-project.github.io/manatee/getting-started/tutorials/

Demo Scenario

- Dataset
  - Insurance charge dataset from Kaggle
- Task
  - Train a model predicting the insurance charge
  - XGBoost Regression
- Privacy Protection
  - Differentially-private synthetic data in the first stage
  - MST (2018 NIST synthetic data challenge winner), McKenna et al.



Name	Modified
backup	4m ago
lost+found	15d ago
sdk	15d ago

Project Current & Future

	Current	Future
Users	One-Way Collaboration (Singler source of data)	Multi-Way Collaboration (Cross-organizational data)
Backend	Single Backend	Multiple Backend
Data Provisioning	Manual	Automated
Policy and Attestation	Manual	Automated
Compute	CPU	CPU/GPU

Project Timeline

- [2024/5] TikTok launched VCE
- [2024/6] TikTok Open sourced PrivacyGo Data Clean Room
- [2024/10] Project renamed to ManaTEE and donated to Confidential Computing Consortium
- [2025/1] ManaTEE community version released
- [Current] Forming Technical Steering Committee

The First Community Release for Open Collaboration

- Tutorial
  - https://manatee-project.github.io/manatee/getting-started/tutorials/
- Local Deployment (Minikube)
  - https://manatee-project.github.io/manatee/getting-started/minikube/
- Announcement
  - https://manatee-project.github.io/manatee/blog/2025/01/07/first-community-release-of-manatee
- Release Note
  - https://github.com/manatee-project/manatee/releases/tag/0.1.0

Collaborators

TikTok

Google

Please Join Us!

Google Groups: https://groups.google.com/u/1/g/manatee-project
Github: https://github.com/manatee-project

Q&A

dayeol.lee@tiktok.com