Designing an Open-Source Platform for Differentially Private Analytics That Is Usable, Scalable, and Extensible

Michael Hay
Tumult Labs & Colgate University
@michaelghay

Presentation at PEPR, June 24, 2022

Tumult **Platform**

- A system for safely releasing aggregate information from sensitive datasets
- Supports standard transformations (filter, join, map, ...) and aggregations (count, average, quantiles, ...)
- Currently used by:

IRS

WIKIMEDIA FOUNDATION

United States Census Bureau

An easy-to-use API capable of powering real-world use cases…

An easy-to-use API capable of powering real-world use cases…

… with provable guarantees of differential privacy…

An easy-to-use API capable of powering real-world use cases…

… with provable guarantees of differential privacy…

… that is extensible…

An easy-to-use API capable of powering real-world use cases…

… with provable guarantees of differential privacy…

… that is extensible…

… and can scale to arbitrarily-sized datasets.

Tumult **Analytics**

Tumult **Analytics**

Intended for data scientists

- DP expertise *not* required
- Python interface similar to pandas/spark

```python
session = Session.from_dataframe(
    dataframe=private_data,
    source_id="my_data",
    PureDPBudget(1.5)
)

query = (
    QueryBuilder("my_data")
    .filter("age > 42")
    .groupby(zip_codes)
    .median("income", low=0, high=10**6)
)

result = session.evaluate(
    query,
    PureDPBudget(0.2)
)

print(session.remaining_privacy_budget())
# prints PureDPBudget(1.3)
```

Tumult **Analytics**

Intended for data scientists

- DP expertise *not* required
- Python interface similar to pandas/spark

User need	Features required
Maximal accuracy	• tight privacy loss accounting • zero-concentrated DP • generalized parallel composition • generalized stability calculus
Data-adaptive workflows	• interactive mechanisms • interactive parallel composition
“Complex” data transformations	• Flatmaps, joins, count distinct, • Bounding user contributions

An easy-to-use API capable of powering real-world use cases…

Tumult Analytics

… with provable guarantees of differential privacy…

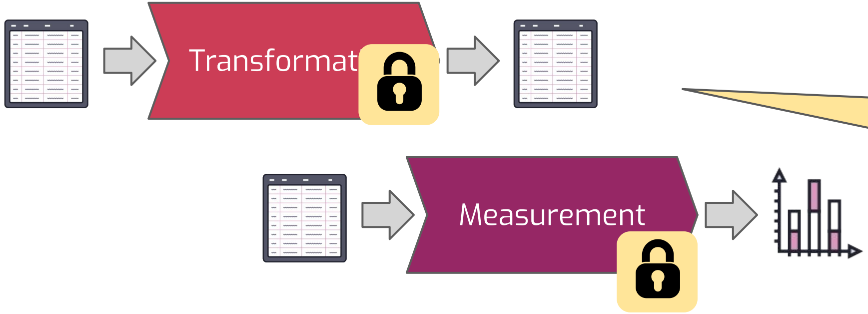
… that is extensible…

Tumult Core

… and can scale to arbitrarily-sized datasets.

Tumult **Core**

A collection of composable components:
transformations and measurements.



Privacy properties

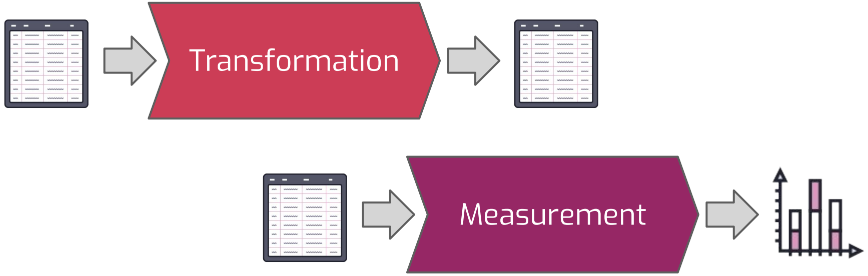
- Input domain
- Input metric
- **Output domain**
- Output **metric**
- **Stability**

Privacy properties

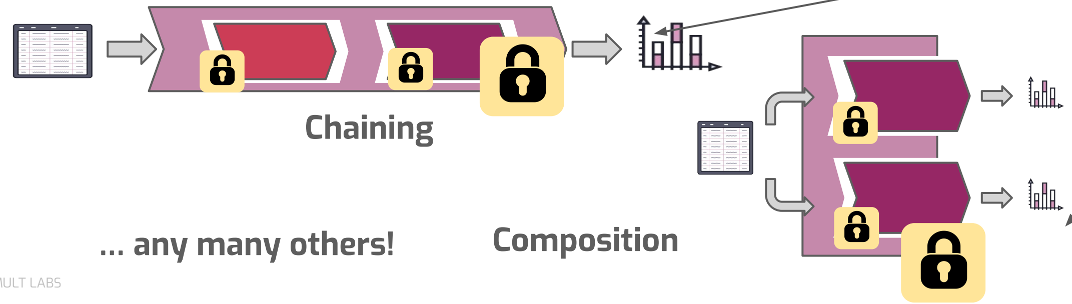
- Input domain
- Input metric
- Output **measure**
- **Privacy loss**

Tumult **Core**

A collection of composable components: transformations and measurements.



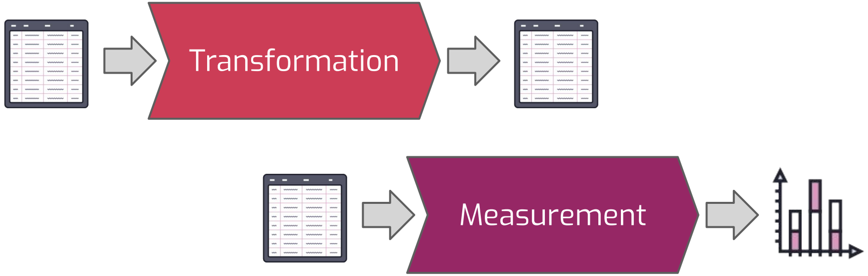
Combinators create new components from existing ones.



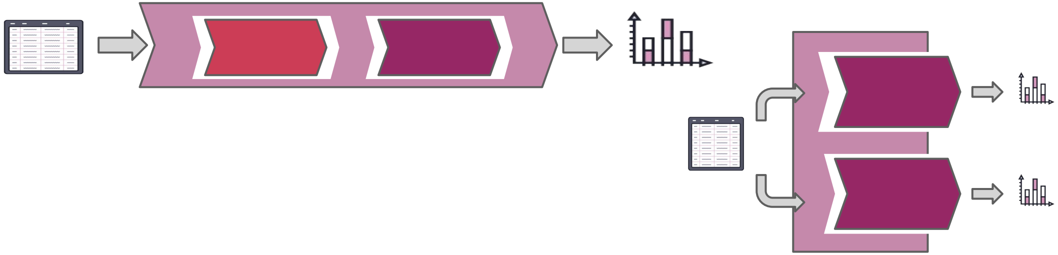
Privacy properties are derived inductively.

Tumult **Core**

A collection of composable components: transformations and measurements.



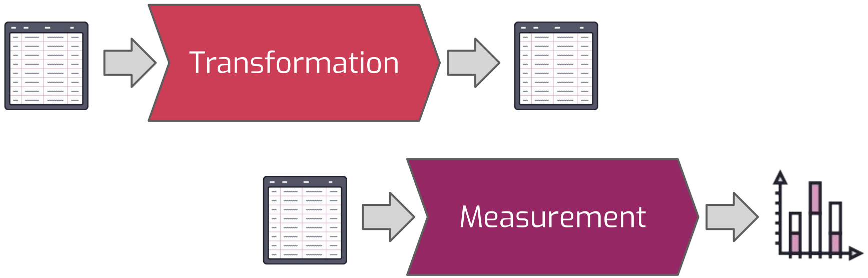
Combinators create new components from existing ones.



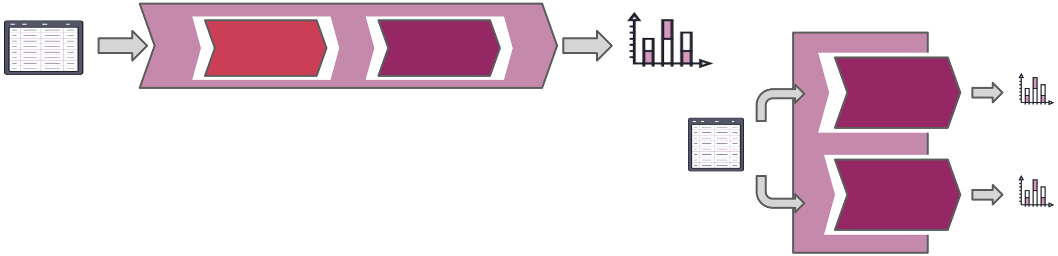
Includes floating-point safe implementations of critical mechanisms (Haney et al. TPDP 2022)

Tumult **Core**

A collection of composable components: transformations and measurements.



Combinators create new components from existing ones.



Core enables creation of complex algorithms from building blocks

Everything carries an explicit, inspectable privacy guarantee

Architecture is modular, extensible

An easy-to-use API capable of powering real-world use cases…

… with provable guarantees of differential privacy…

… that is extensible…

… and can scale to arbitrarily-sized datasets.

Tumult **Analytics**

Tumult **Core**

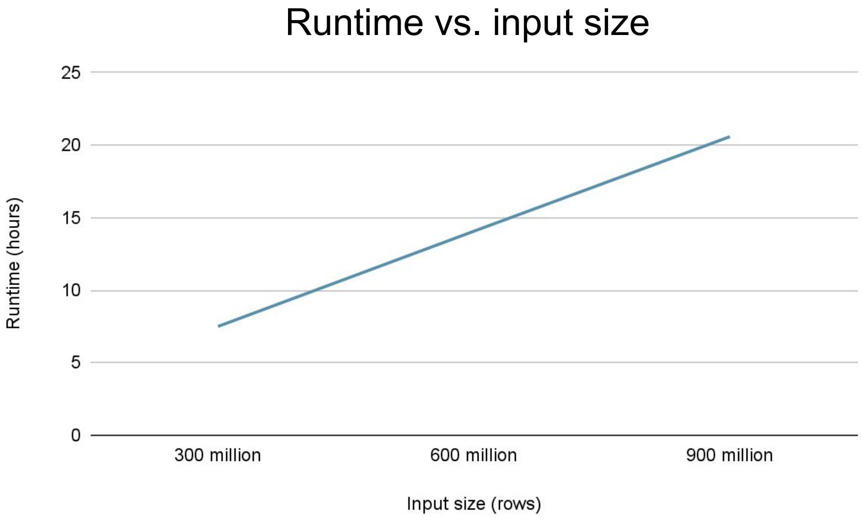
Scalability Experiments

Simulations on synthetic dataset

Input rows: age, sex, detailed race, geography

Output: histograms (age x sex) for each race group at varying levels of geography

Input size (rows)	Input size (GB)	Output size (rows)	Output size (GB)	Runtime
300 million	41.3	62 million	1.30	7h 30m
600 million	82.7	64 million	1.37	14h 4m
900 million	115.0	65 million	1.39	20h 35m



Experiments were run on r4.8xlarge (1 master, 2 worker nodes)

An easy-to-use API capable of powering real-world use cases…

… with provable guarantees of differential privacy…

… that is extensible…

… and can scale to arbitrarily-sized datasets.

Tumult **Analytics**

Tumult **Core**

APACHE Spark™

🎉 Open-source release: today 🎉

gitlab.com/tumult-labs/analytics

gitlab.com/tumult-labs/core

We'd love your feedback!
(See slack for details)

Thank you!

Michael Hay
michael@tmlt.io
@michaelghay

gitlab.com/tumult-labs/analytics
gitlab.com/tumult-labs/core

tmlt.io/connect
tmlt.io/careers
@TumultLabs