



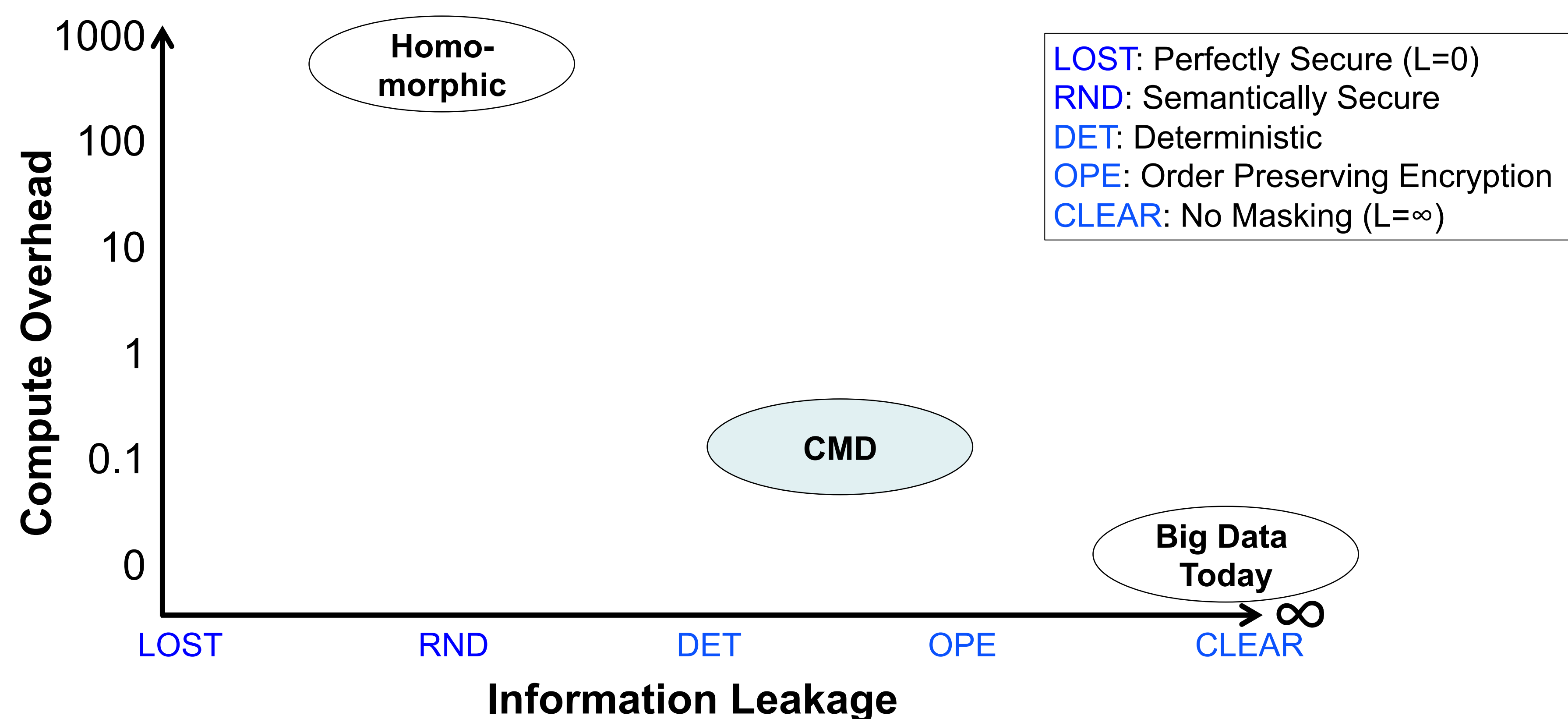
Common Big Data Architecture

The diagram illustrates the layers of a big data architecture and the various attack vectors that can target it:

- Users:** The top layer of the architecture.
- Operators:** A layer below Users, containing an **External Denial Of Service** attack vector.
- Analysts:** A layer below Operators, containing a **Credential Stealing** attack vector.
- Commanders:** A layer below Analysts.
- Web:** A layer below Commanders, containing a **Remote Code Injection** attack vector.
- Databases:** A layer below Web, containing a **Data Loss / Exfiltration** attack vector.
- Analytics:** A layer below Databases, containing an **Insider Threat** attack vector.
- Files:** A layer below Analytics, containing a **Data Integrity Attack** attack vector.
- Ingest:** A layer below Files, containing a **Data Integrity Attack** attack vector.
- Scheduler:** A layer below Ingest, containing a **Hypervisor Privilege Escalation** attack vector.
- Computer:** A layer below Scheduler, containing an **Internal Network Resource Attacks** attack vector.
- Supply Chain:** A layer below Computer, containing a **Supply Chain** attack vector.
- Data:** The bottom layer of the architecture.

Red arrows indicate the flow of data and the specific attack vectors targeting each layer. The diagram also shows data flow between components like Files, Ingest, Databases, and Analytics.

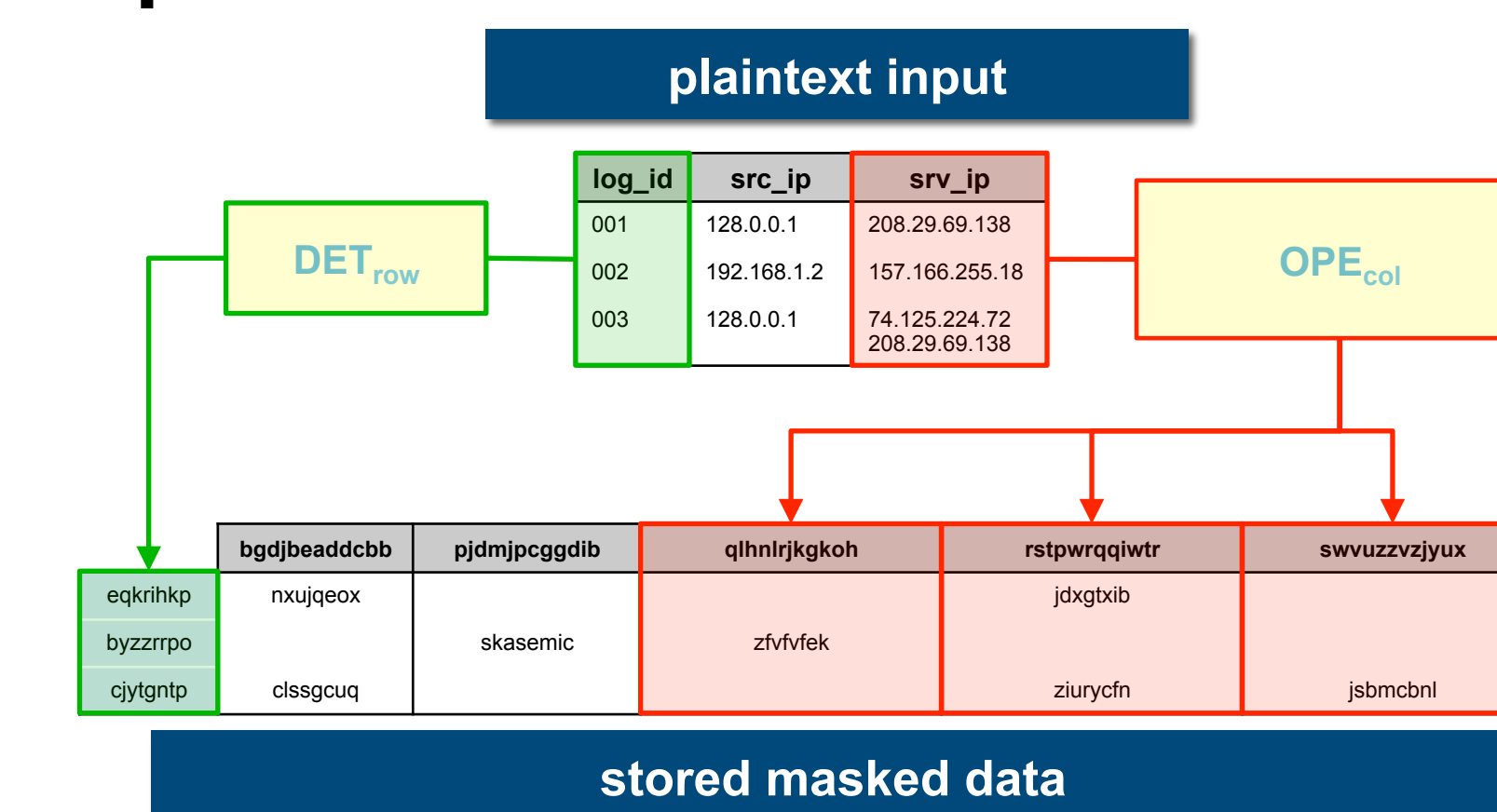
Performance Trade Space



- **Computing on masked data (CMD)** raises the bar on data in the clear
- **Uses lower over head approaches** such as deterministic (DET) encryption and order preserving encryption (OPE)
- **Associative array (D4M) algebra** is defined over sets (not real numbers); allows linear algebra to work on DET or OPE data

- Associative array algebra defined on strict totally ordered sets
- Algebra works with DET or OPE on rows or columns by definition
- Values can often be operated on with RND or LOST

Step 1: Mask Data in Database



Step 2: Mask & Compute Analytic

Pedigreed correlation calculation in plain text

```
A = CatKeyMul(      T(:, 'src_ip|000. : src_ip|999.').',
                    T(:, 'srv_ip|000. : srv_ip|999.')) )
```

Pedigreed correlation calculation in masked text

```
A = CatKeyMul(      T(:, 'bgdjbeaddcbb : pjdmjpcggdib '),
                    T(:, 'qlhnlrjkqkoh : swvuzzvzjyux ') )
```

Tallies row IDs between vertices

A	qlhnlrjkgkoh	rstpwrqqlwtr	swvuzzvzyux
bgdjbeaddcbb		eqkrihkp cyjtgntp	cyjtgntp
pjdmjpcggdib	cyjtgntp		

Step 3: Unmask

A	srv_ip 157.166.255.18	srv_ip 208.29.69.138	srv_ip 74.125.224.72
src_ip 128.0.0.1		log_id 100 log_id 300	log_id 300
src_ip 192.168.1.2	log_id 200		

- Analytic computed on database or another server
- Masked output is sent back to user
- D4M client unmaskes data

This work is sponsored by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.