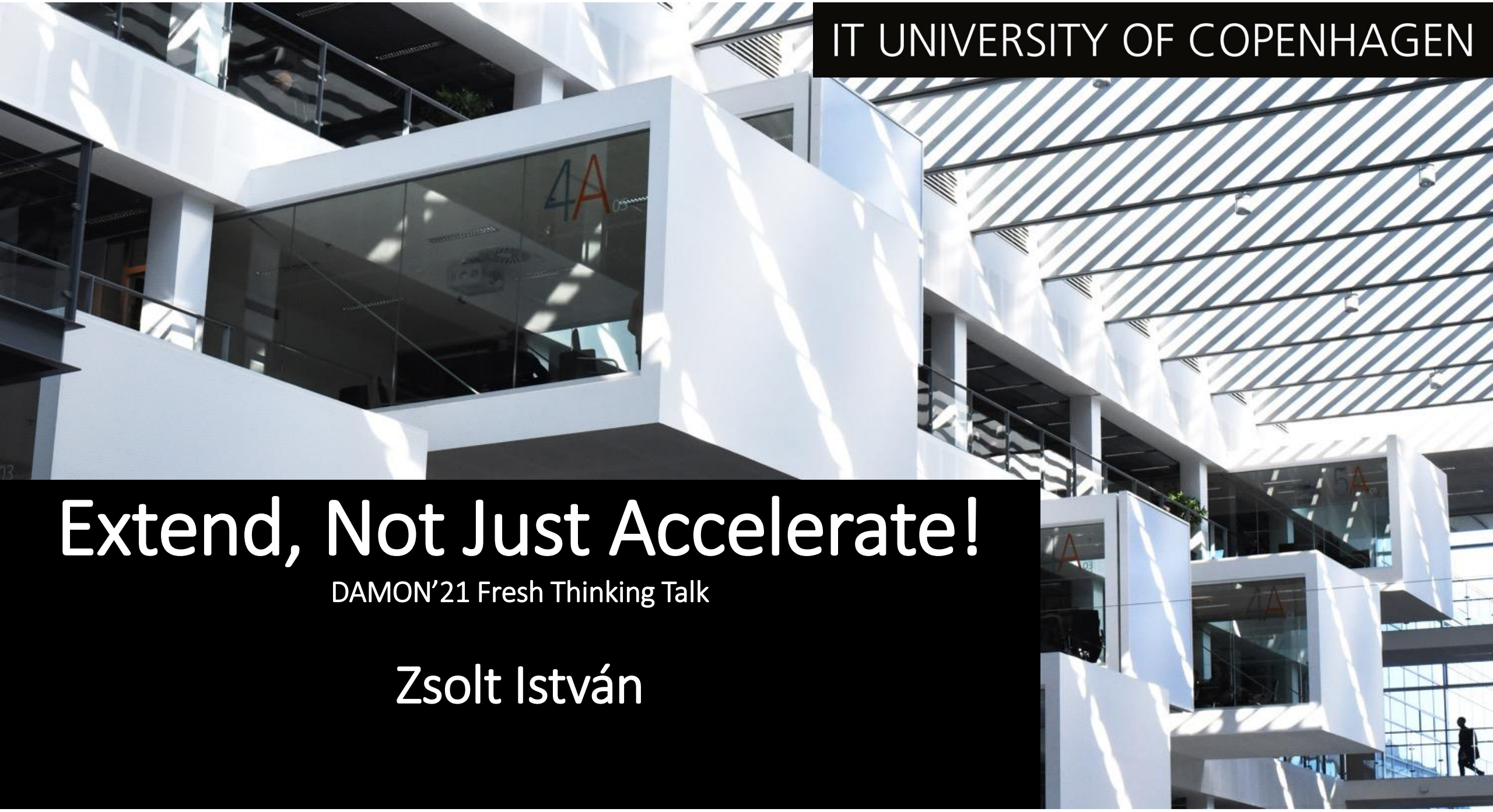
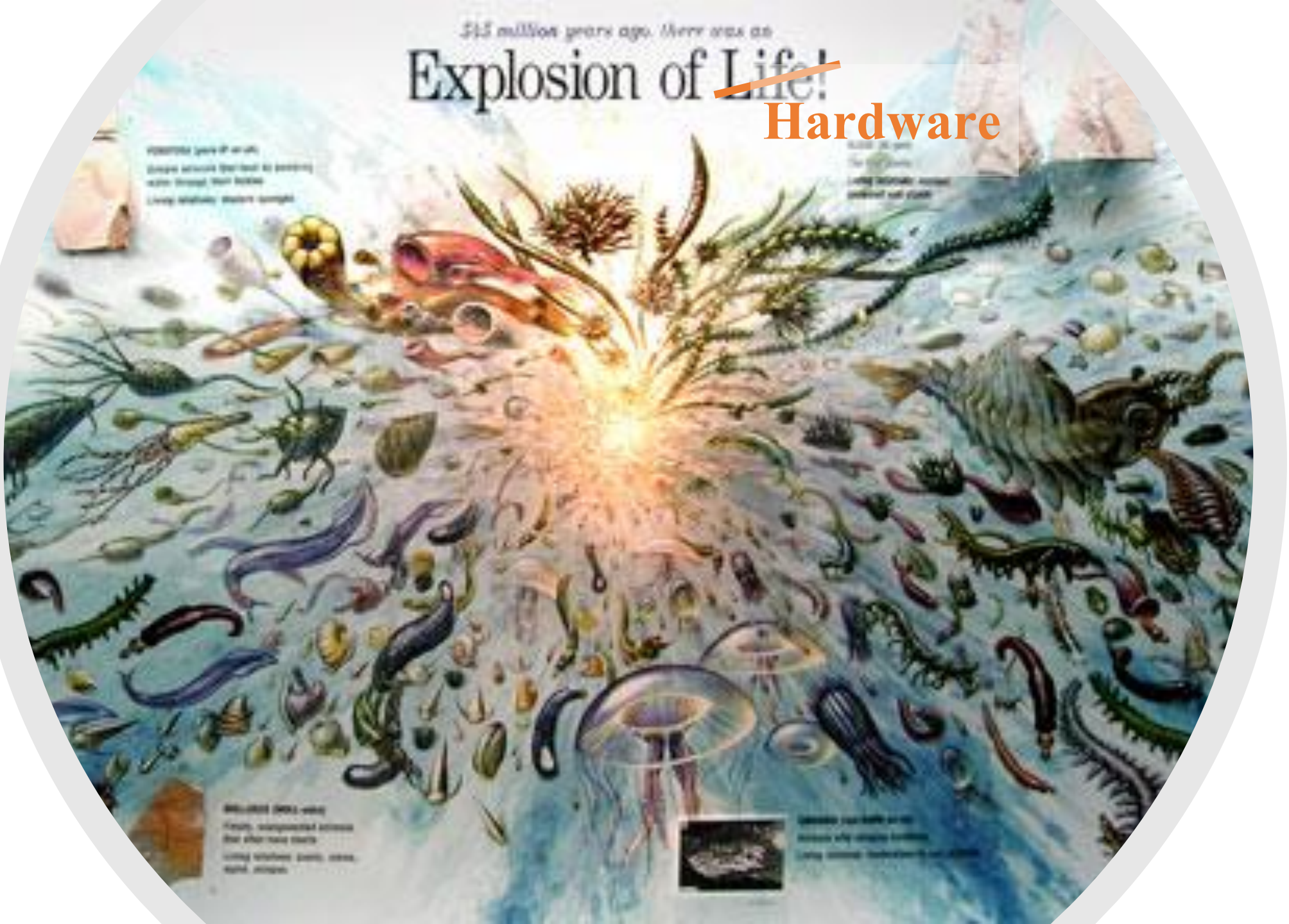


Extend, Not Just Accelerate!

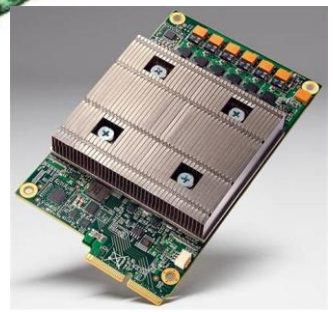
DAMON'21 Fresh Thinking Talk

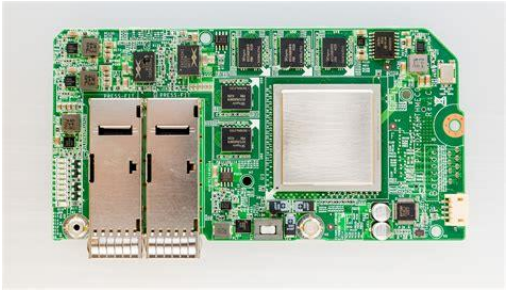
Zsolt István


Explosion of Life!
Hardware

Specialized Hardware is Everywhere

- Accelerators (GPGPUs, TPUs, FPGAs, …)
- Fast and Smart Networking (NICs, Switches, …)
- Smart Drives, ZNS, …
- Motherboards with spec. chips
- …

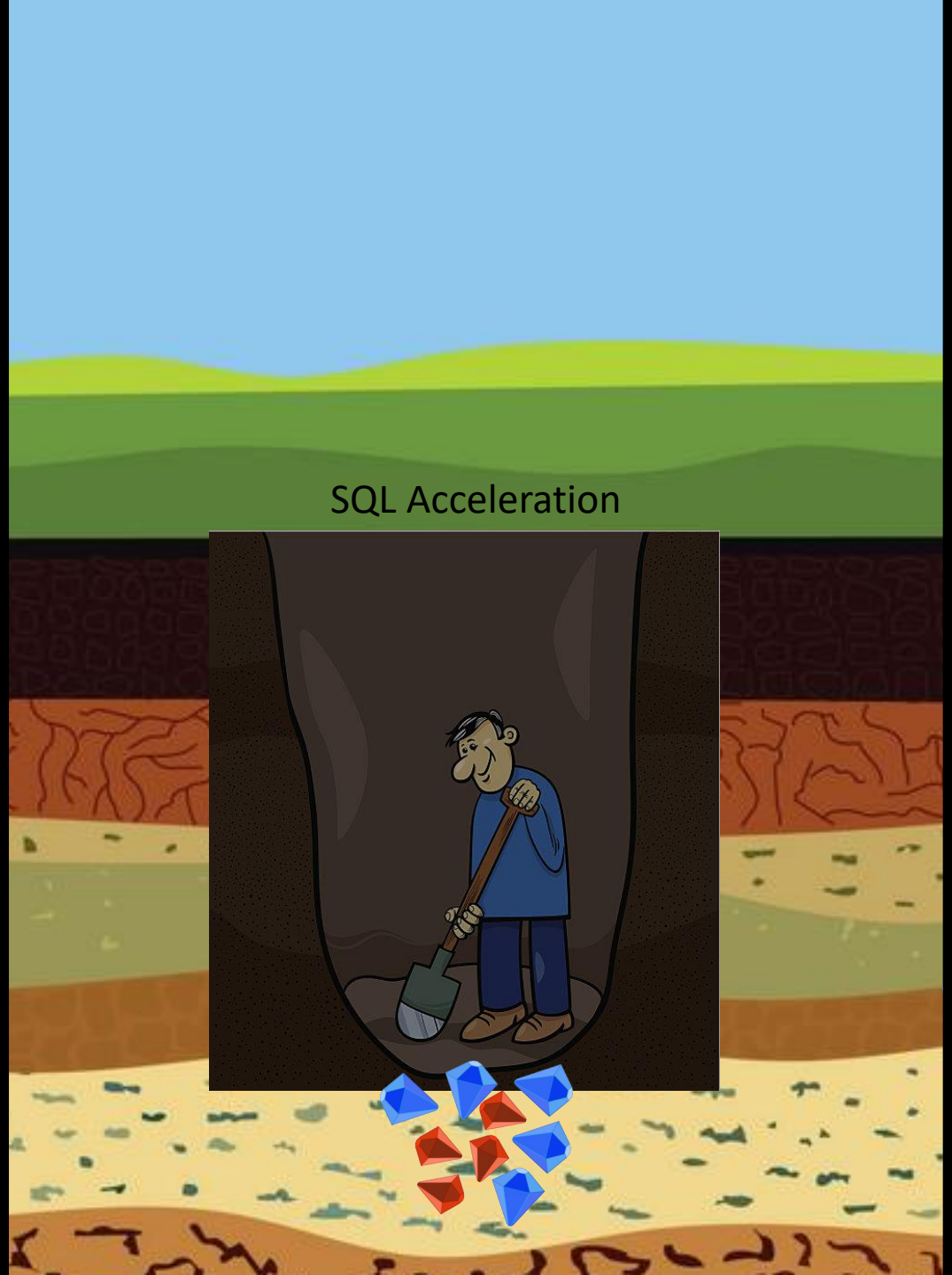


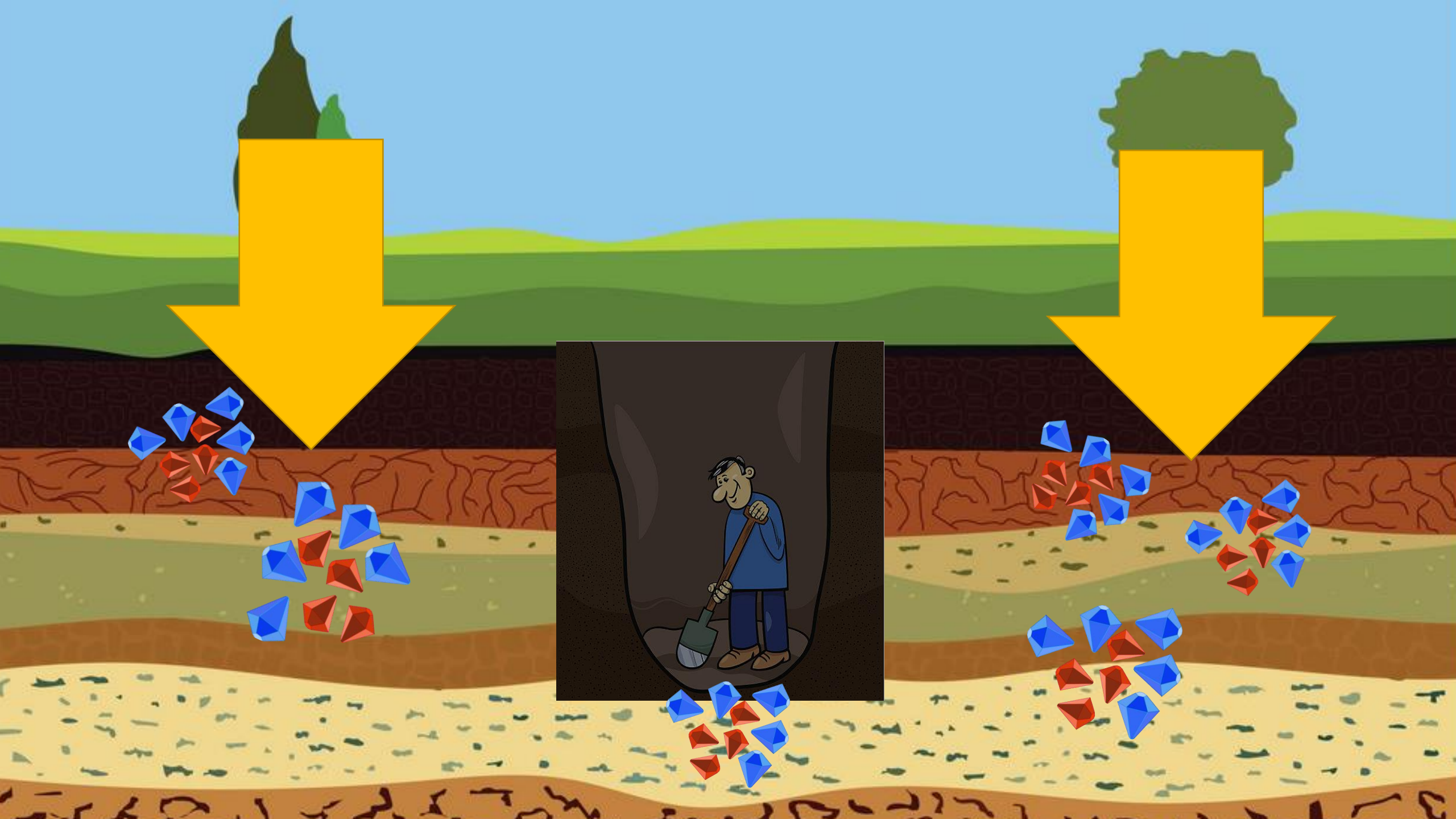


NITRO CARDS

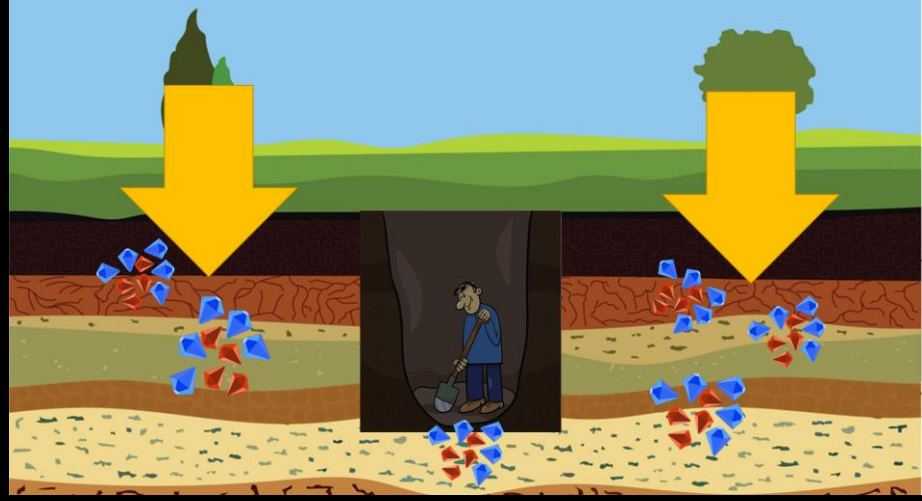
Use to Improve Databases

SQL Acceleration





Extend, Not Just Accelerate!



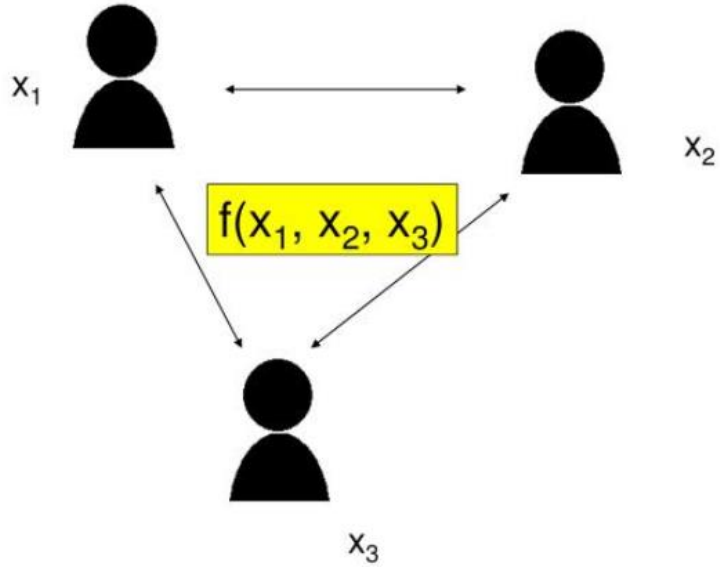
- Add new features or guarantees to DB
- Utilize modern hardware to hide the cost, or to make practical

- This talk: A sample of exciting areas to look at as Database/Systems person
  (Non-exhaustive list of related work)

Privacy Preserving Operators

Secure Multiparty Computation (MPC)

- Secure Multiparty Computation
  - Participants compute a shared function without disclosing their part of the data



- Many database applications: Anything that would need shared data!
  - E.g.: statistics on sicknesses based on patients with records at different hospitals

Challenges in MPC

- Orders of magnitude slower than plaintext processing
  - Computation expressed as “circuit”
  - “Gates” evaluated with cryptographic functions

- Most optimizations today are at the algorithm level, next step will be HW!

  [Volgushev et al. Conclave: secure multi-party computation on big data. EuroSys’19]

- There are more and more libraries to work on:

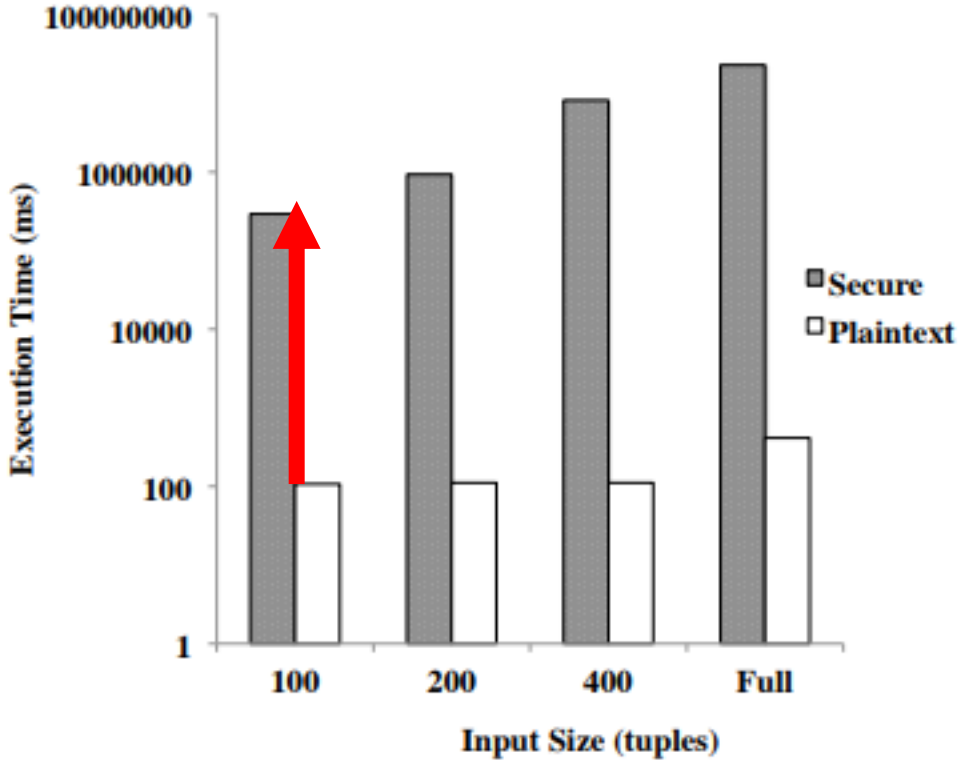
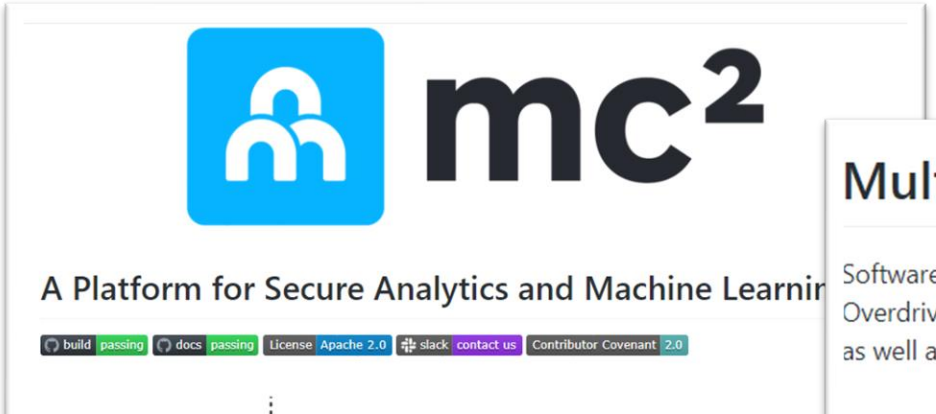
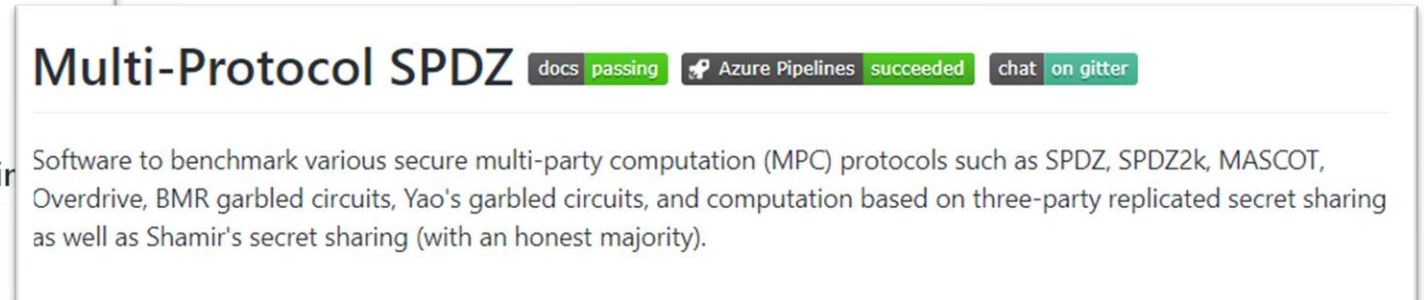


Figure 9: Runtime of *comorbidity* on increasing data sizes.

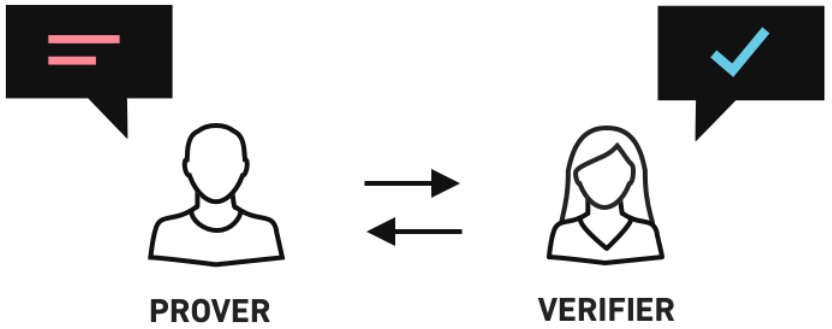
[Bater, Johes, et al. "SMCQL: Secure Query Processing for Private Data Networks." Proc. VLDB Endow. 10.6 (2017): 673-684.]





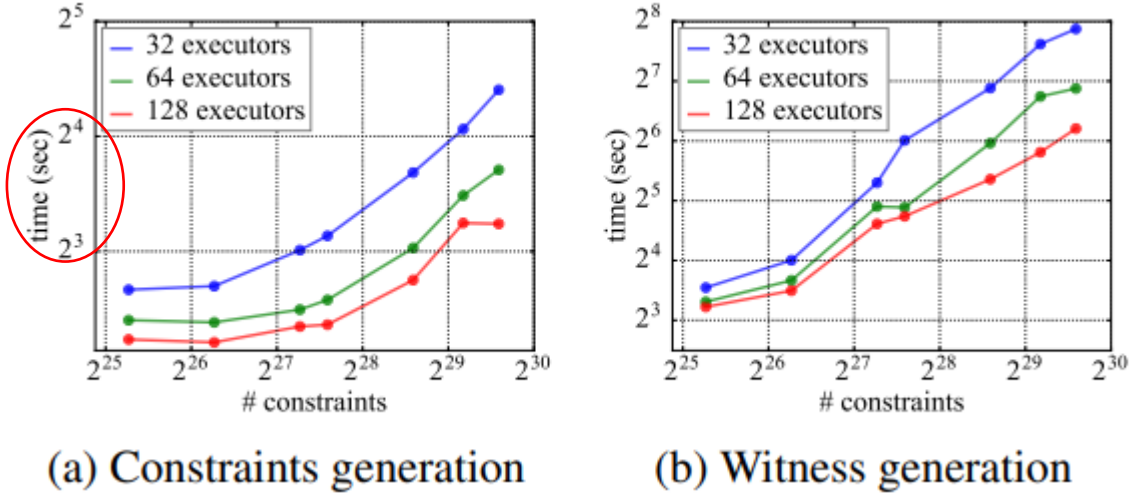
Zero Knowledge Proofs (ZKP)

- One party proves to another they know *x* without disclosing it
  - Used in privacy-preserving cryptocurrency like Zcash
  - Opportunities: Auditing/Regulations in Databases



- Notoriously slow to prepare proof (seconds!) but verification is cheap (ms)
  - Operations on very large polynomials, significant setup overhead

ZKP Challenges



- Distributed computation can speed it up (distributed matrix operations on a large matrix)

  [Wu et al. DIZK: A Distributed Zero Knowledge Proof System. USENIX Security 2018]

- Hardware ideas appearing to specialize for underlying operations

  [PipeZK: Accelerating Zero-Knowledge Proof with a Pipelined Architecture. ISCA 2021]

- Integration and optimization for DBs?

Policy Compliance

Policy Compliance in Databases

- Decades work on access control, lineage, provenance, etc.
  - Now we have a legal requirement!

- Many emerging works on aiming to ensure, e.g., CCPA or GDPR compliance

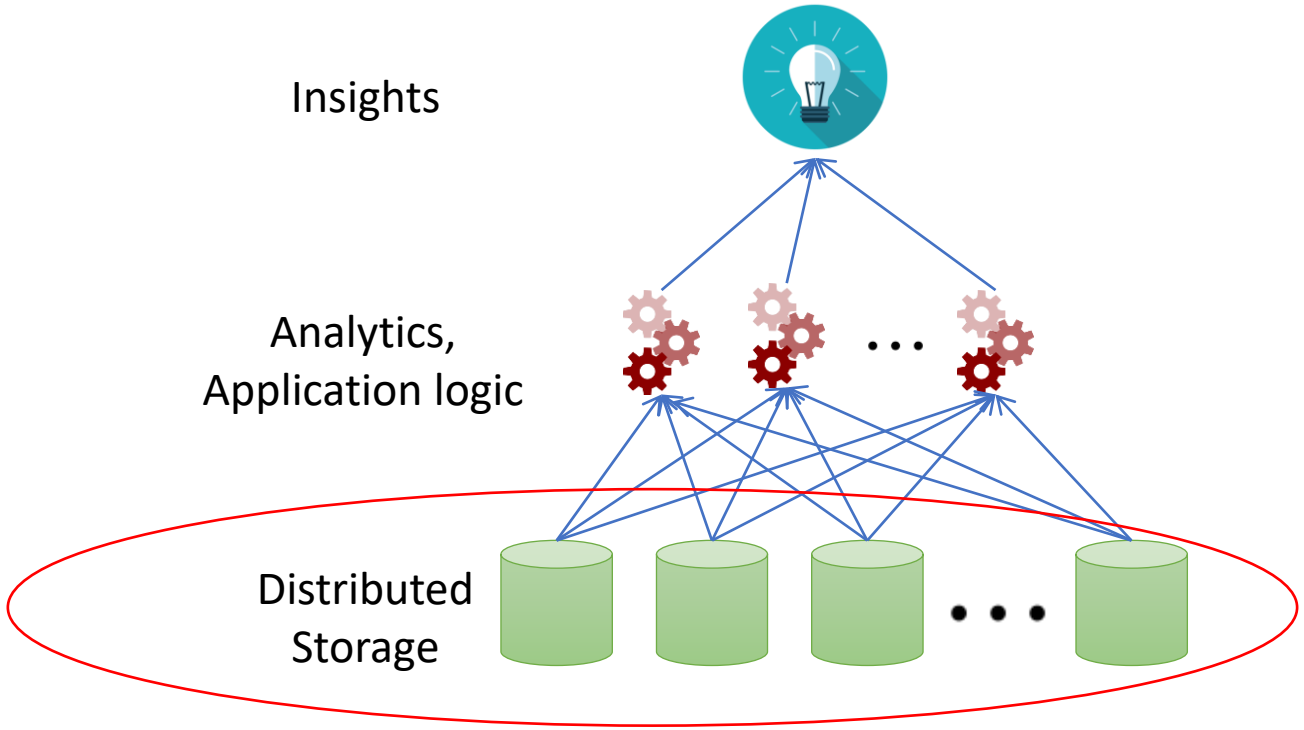
  [Kraska et al. SchengenDB: A Data Protection Database Proposal, Poly/DMAH@VLDB 2019]

  [Marzoev et al. Towards Multiverse Databases. HotOS 2019]

  [Shashtri et al. Understanding and Benchmarking the Impact of GDPR on Database Systems. VLDB20],…

There is an associated slowdown!

Disaggregated Architectures and Compliance



Software-defined Data Protection – VLDB Vision with


Soujanya Ponnapalli,
Vijay Chidambaram

How GDPR Affects Storage

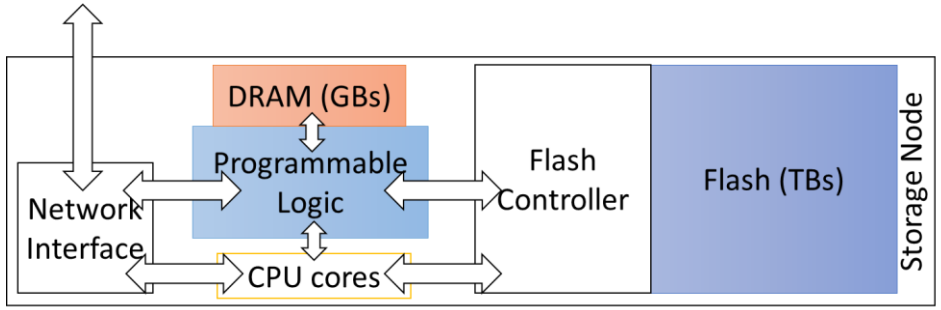
- GDPR: >30% of data protection articles affect storage

  [Shah et al. Analyzing the Impact of GDPR on Storage Systems. HotStorage 19]

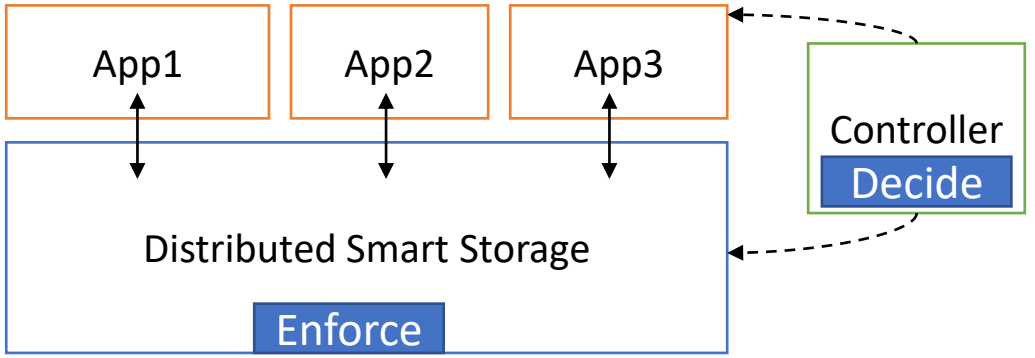
No.	GDPR article	Required functionality
5.1	Purpose limitation (data collected for specific purpose)	Fine-grained permissions
21	Right to object (data not used for objected reason)	Fine-grained permissions
5.1	Storage limitation (data not stored beyond purpose)	Deletion
17	Right to be forgotten	Deletion
15	Right of access by users	Metadata (and Secondary indexes)
20	Right to portability (transfer data on request)	Metadata (and Secondary indexes)
5.2	Accountability (ability to demonstrate compliance)	Logging and Monitoring
30	Records of processing activity	Logging
33, 34	Notify data breaches	Logging and Monitoring
25	Protection by design and by default	Encryption
32	Security of data	Encryption and Access control
13	Obtain user consent on data management	High level policy†
46	Transfers subject to safeguards	Location control†

Using Heterogenous Hardware

- Emerging Smart Storage nodes with heterogenous compute
  - Can levy re-imagined processing
  - Policies require complex decision making…



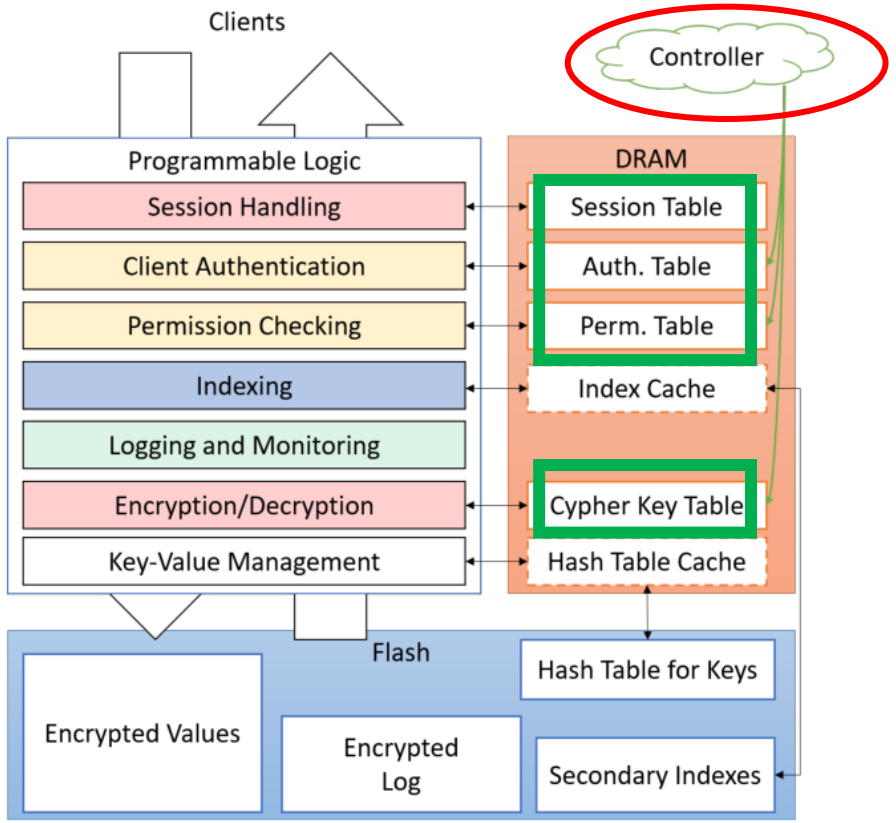
- **Software-Defined Data Protection (SDP)**
  - Decoupling enforcement from decisions increases performance



The design allows for complying with complex rules at network line-rate, e.g., with GDPR!

SDP Pipeline

No.	GDPR article	Required functionality
5.1	Purpose limitation (data collected for specific purpose)	Fine-grained permissions
21	Right to object (data not used for objected reason)	Fine-grained permissions
5.1	Storage limitation (data not stored beyond purpose)	Deletion
17	Right to be forgotten	Deletion
15	Right of access by users	Metadata (and Secondary indexes)
20	Right to portability (transfer data on request)	Metadata (and Secondary indexes)
5.2	Accountability (ability to demonstrate compliance)	Logging and Monitoring
30	Records of processing activity	Logging
33, 34	Notify data breaches	Logging and Monitoring
25	Protection by design and by default	Encryption
32	Security of data	Encryption and Access control
13	Obtain user consent on data management	High level policy†
46	Transfers subject to safeguards	Location control†



- Achievable with state-of-the-art
- Same interface for different HW and controller implementations
- Most remaining challenges in Controller!

SDP Challenges

- Software controller: convert from “laws” to HW rules

  [Krahn et al. Pesos: Policy Enhanced Secure Object Store. EuroSys’18]

  [Upadhyaya et al. Automatic Enforcement of Data Use Policies with DataLawyer. SIGMOD 15], …

- Have to trust Storage Firmware not to leak keys, etc.

Need custom TEEs!

- First step: TEEs with FPGAs in the cloud

  [Zeitouni et al. Trusted Configuration in Cloud FPGAs. FCCM 2021]

  [Zhao et al. ShEF: Shielded Enclaves for Cloud FPGAs. Arxiv]

- How to partition work between SW/HW?

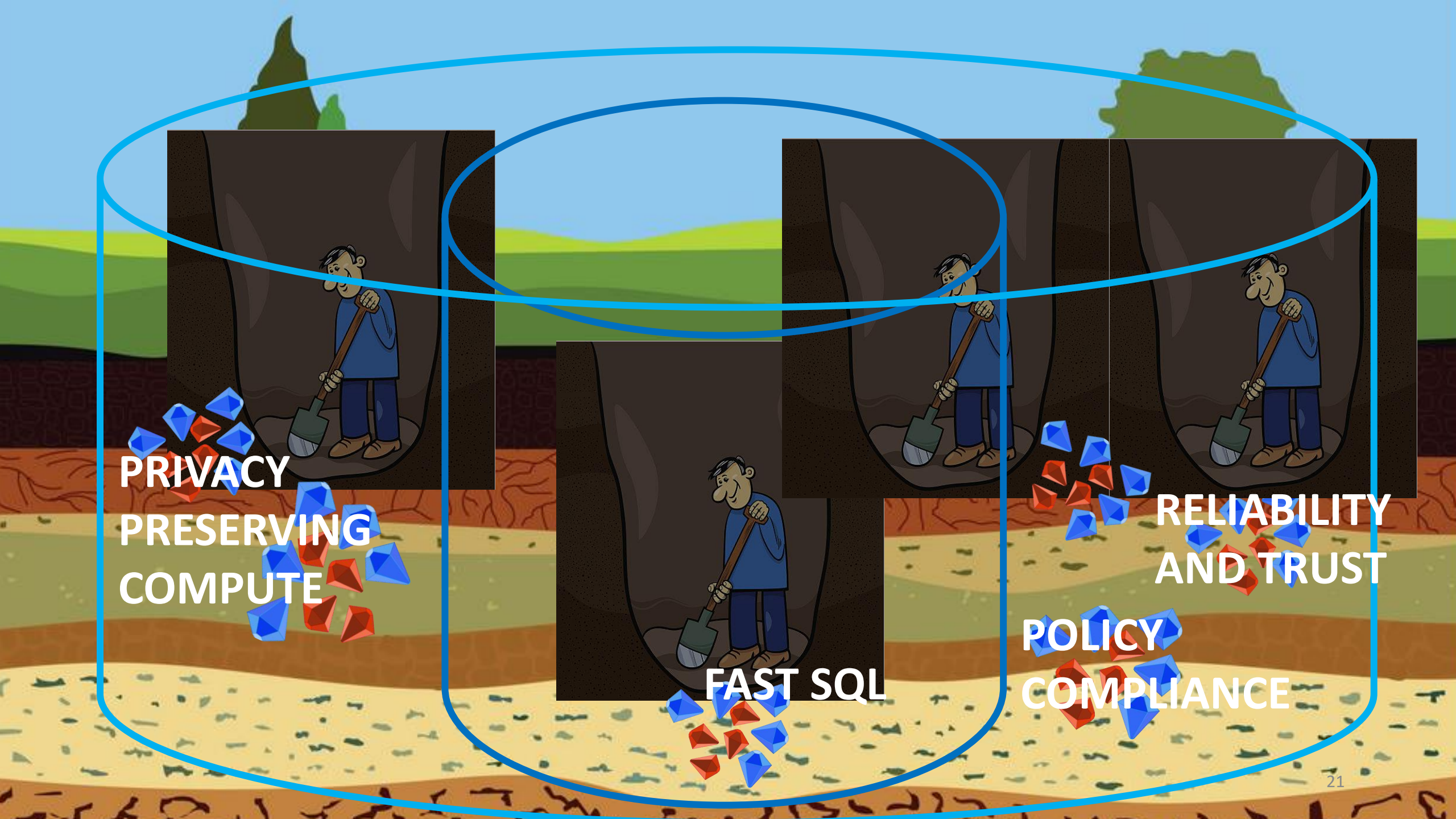
Reliability and Trust

Reliability and Trust

- Databases/Analytics are becoming massively distributed
- Techniques from Blockchains can be useful:
  - Decentralization of trust
  - Byzantine Fault Tolerant consensus

- Slower than “regular” distributed algorithms – compute intensive *and* data movement intensive
  - Hardware can help!

Experimental Paper at DEBS’21 with Man-Kit Sit and Manuel Bravo


PRIVACY
PRESERVING
COMPUTE
FAST SQL
RELIABILITY
AND TRUST
POLICY
COMPLIANCE