Getting Rid of Data

Tova Milo
Tel Aviv University

The Big Data Era

From sports,
to health care,
to the way we drive our cars,
or choose how to invest our money,…
Big Data is changing every aspect of our lives.

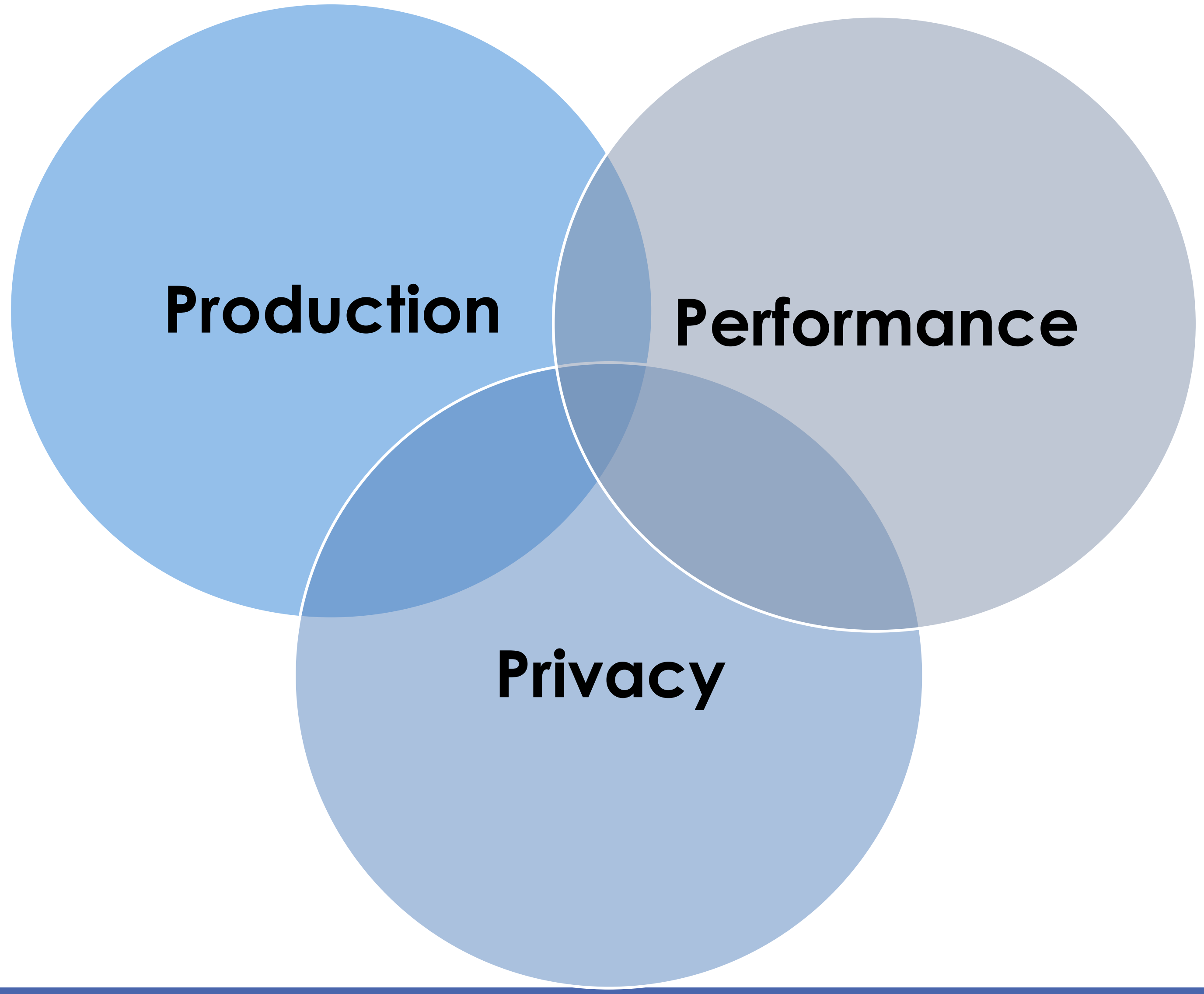
The Big Data Era

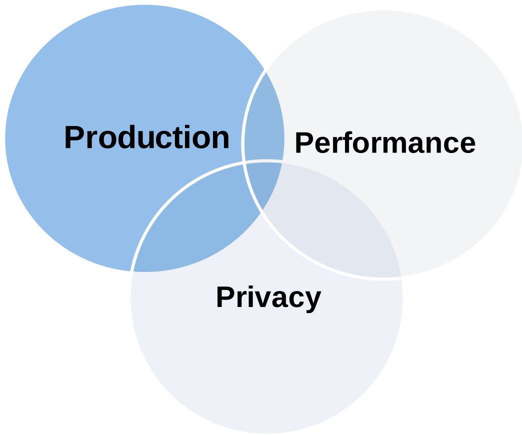
The **data-centered revolution** is fueled by the masses of data, but at the same time is at a great risk due to the very same **information flood**.

The Big Data Era

Time to stop and rethink the “More Data!” philosophy.

The 3 P’s to worry about:





Production of Data & Storage

The size of our digital universe grows exponentially

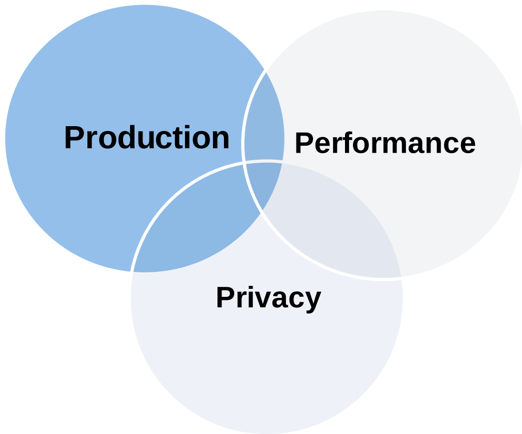
Forecast [IDC'17]:

"By 2025 the global datasphere will grow to **163 zettabytes** (trillion giga), ten times the **16.1 ZB** of data generated in 2016."

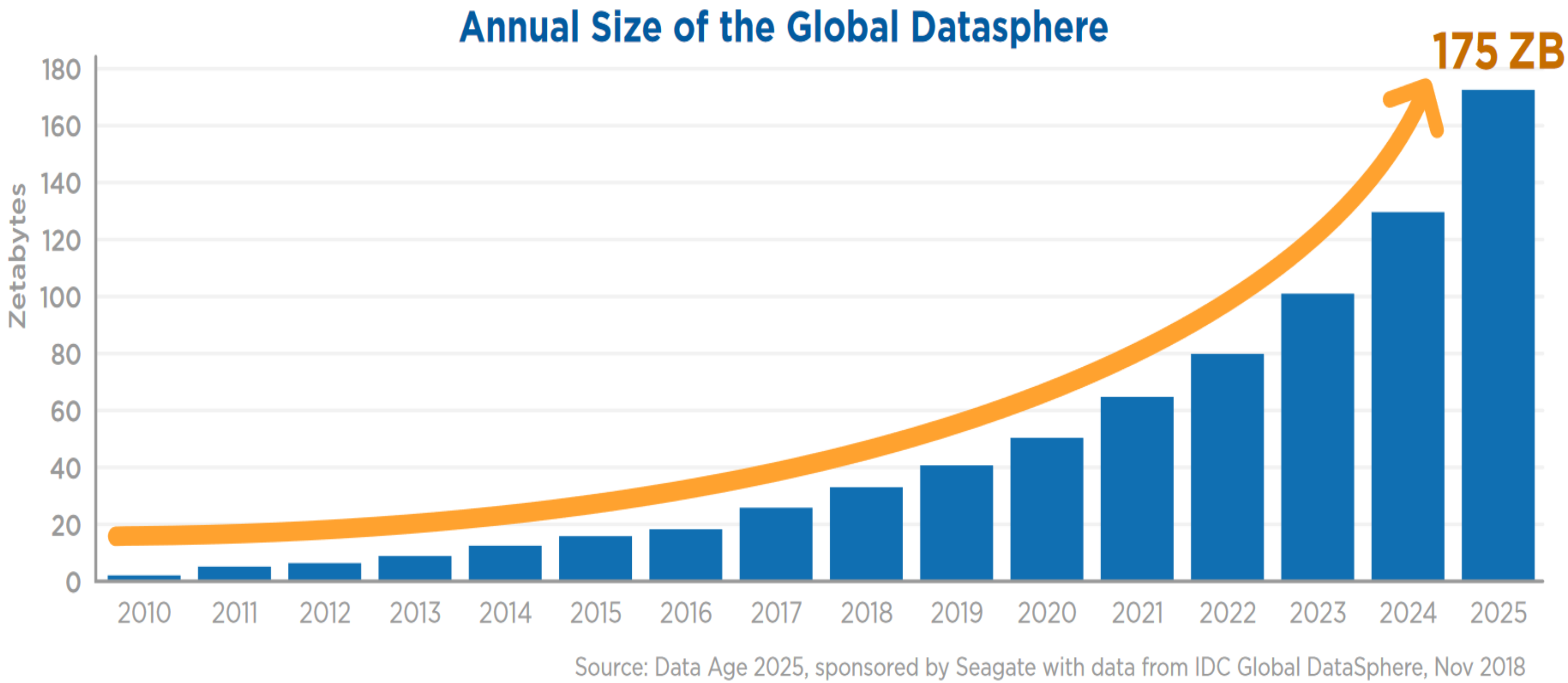
Updated forecast [IDC'18]:

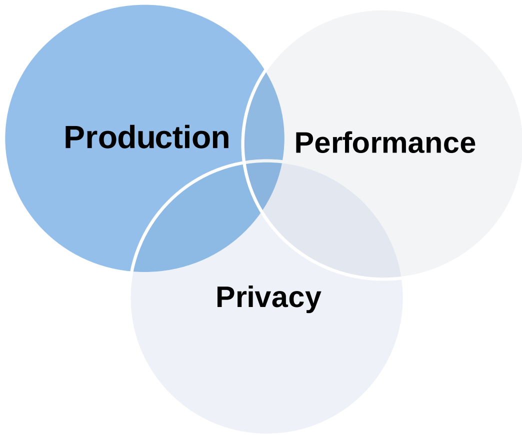
"By 2025 the global datasphere will grow to **175 zettabytes**, from the **33 ZB** in 2018"

Storage demand is estimated to outstrip production by more than double!


Production
Performance
Privacy

Data Size

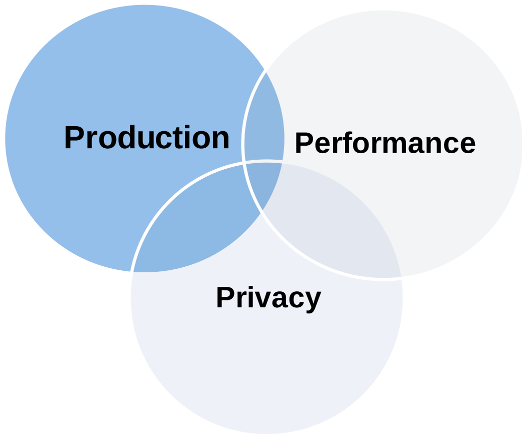

Annual Size of the Global Datasphere
175 ZB
Zetabytes
180
160
140
120
100
80
60
40
20
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018



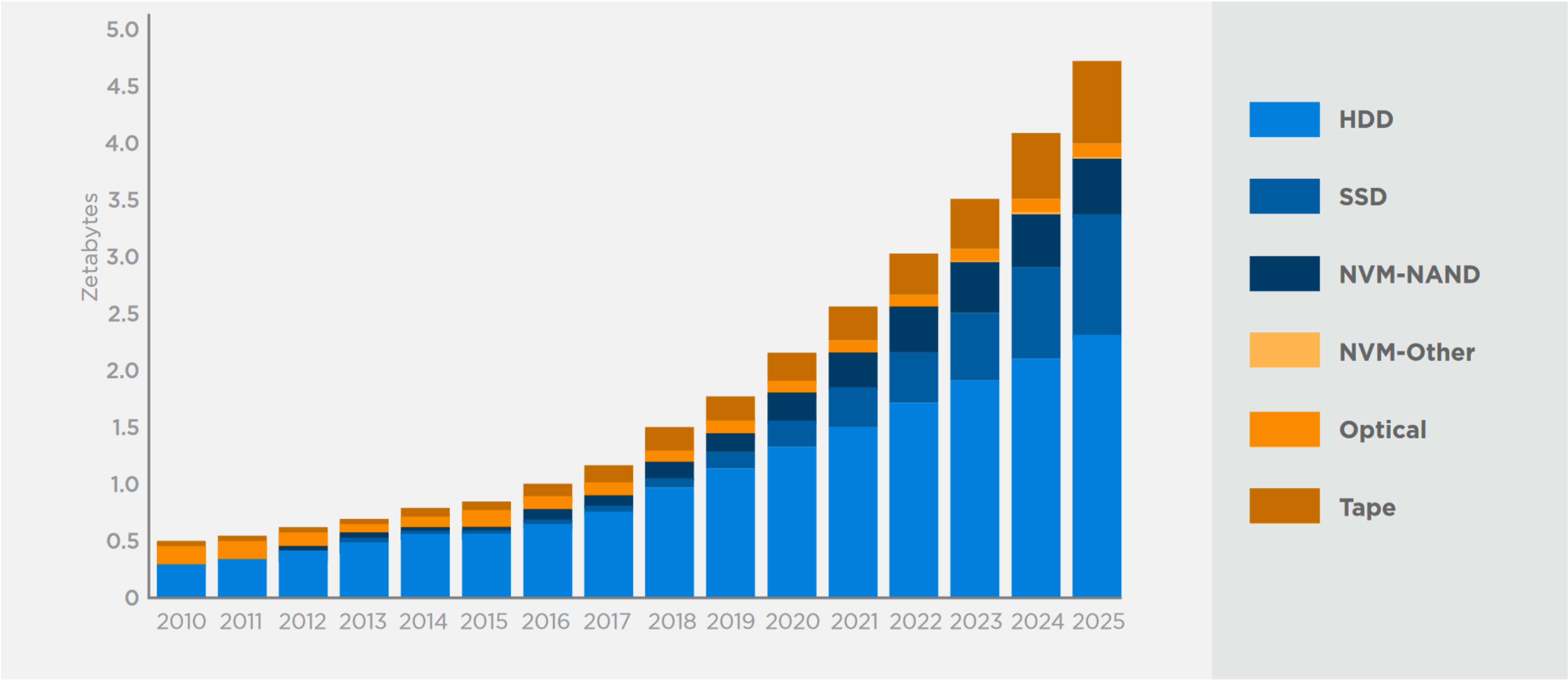
How Much is175 ZB?

> “If one were able to store 175ZB onto BluRay discs, then you’d have a stack of discs that can get you to the moon 23 times…”

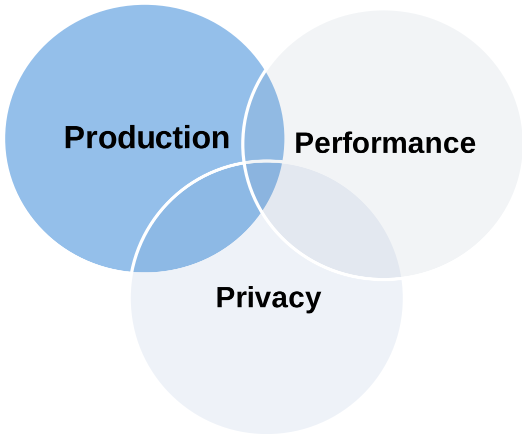
> “Even if you could download 175ZB on today’s largest hard drive it would take 12.5 billion drives (and as an industry, we ship a fraction of that today.)”



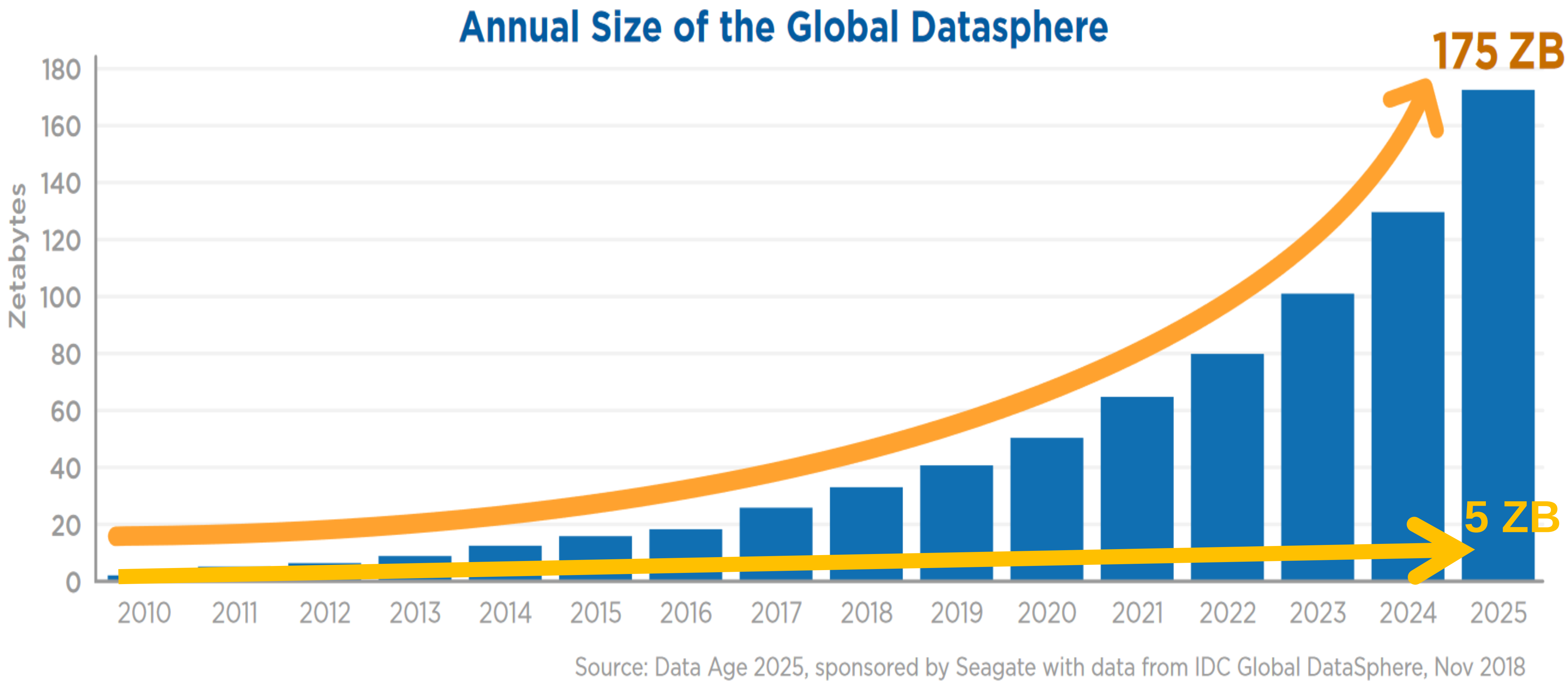
Storage Production

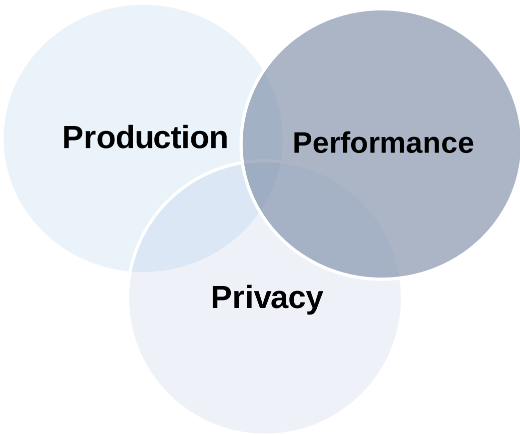


Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018



Data vs. Storage



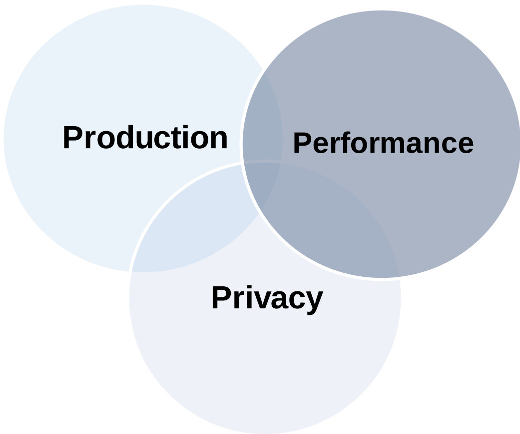


Performance

Handling exponentially growing data incurs a substantial maintenance and processing overhead

- data cleaning,
- validation,
- enhancement,
- analysis,…

Selective data management is key to performance !



Let's Think Energy…

DATA | ECONOMY

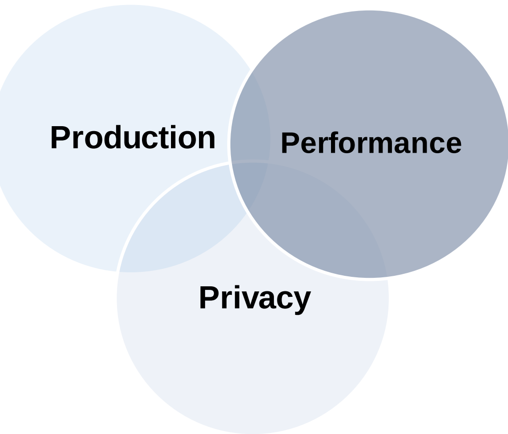
NEWS ▾ ECONOMY ▾ BUSINESS ▾ MARKETS ▾ LEADERSHIP ▾ INDUSTRY ▾ LIFE & AR

TRENDING Salesforce completes $15.7bn acquisition of analytics thoroughbred Tableau Software

Data Centres | World

Data Centres Of The World Will Consume 1/5 Of Earth's Power By 2025

By João Marques Lima | **PUBLISHED:** 05:30, 12 December, 2017 | **UPDATED:** 00:32, 12 December, 2017



Let's Think Energy…

DATA | ECONOMY   NEWS   ECONOMY   BUSINESS   MARKETS   LEADERSHIP   INDUSTRY   LIFE & AR

TRENDING Salesforce completes $15.7bn acquisition of analytics thoroughbred Tableau Software

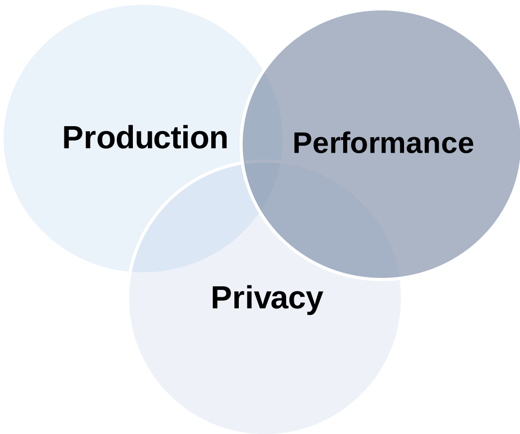
Data Centres   World

Data Centres Of The World Will Consume 1/5 Of Earth's Power By 2025

Globally, data centres were in 2014 responsible for around 1.62% of the world's utilised energy that year, according tc Yole Développement.

That has increased today to more than 3% of the world's energy (around 420 terawatts) and data centres are also responsible for 2% of total greenhouse gas emissions.

More On: Renewable Energy | Power | Green Data Center

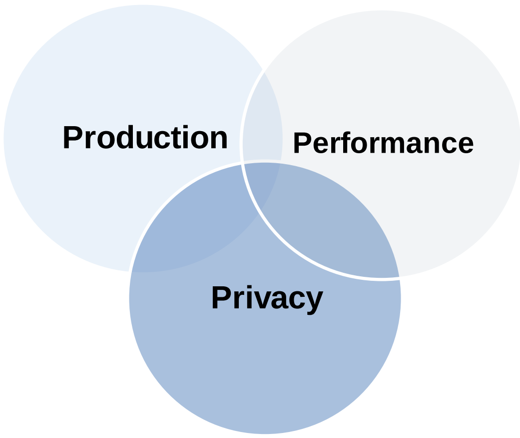


Energy Optimization ?

Over the last few years:

- Development of better ways to cool data centers
- Recycling the waste heat
- Streamlining computing processes
- Switching to renewable energy

Still, even in the best-scenario predictions, if we don't learn how to dispense of data we'll stay at the same consumption level (which is already high)



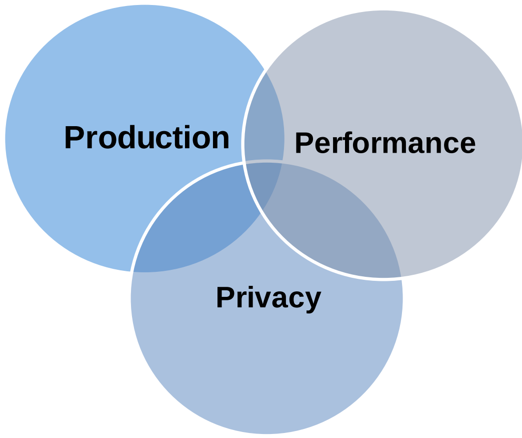
Privacy and Security

Even if we disregard storage and performance constraints, uncontrolled data retention dangers privacy & security

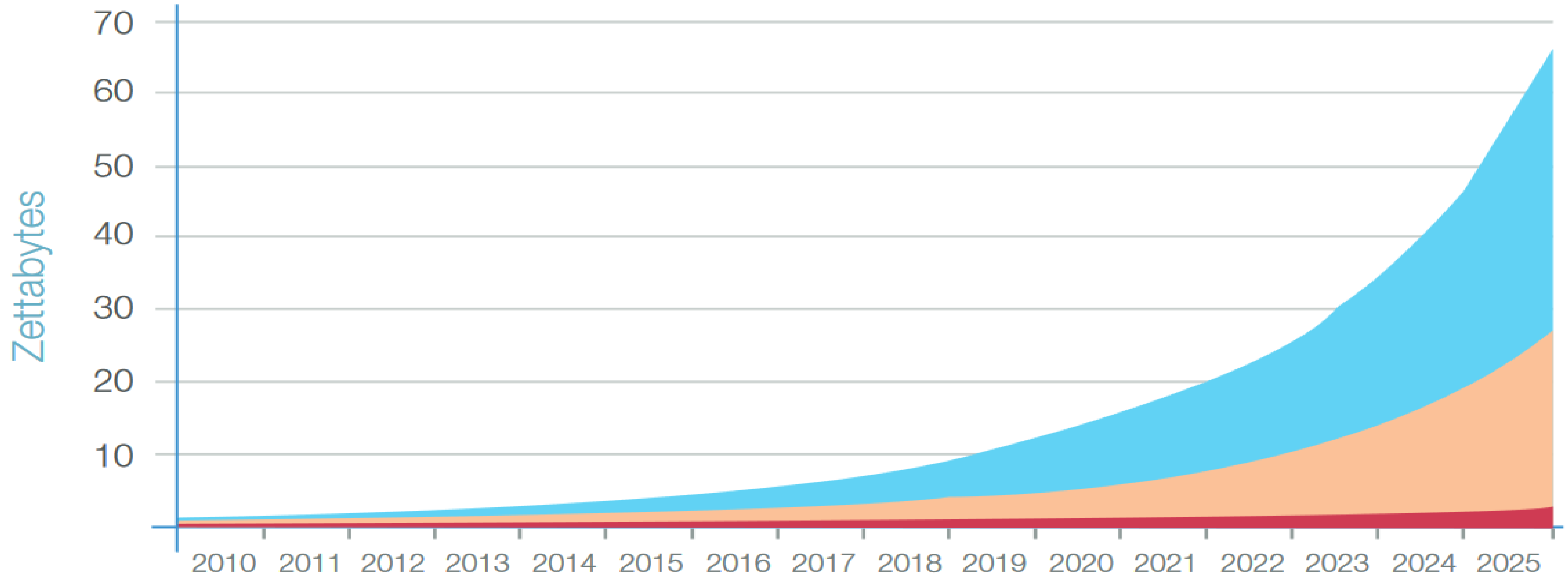
- EU Data Protection Regulation (GDPR).
- Sarbanes-Oxley, Graham-Leach-Bliley, the Fair and Accurate Credit Transactions Act, HIPAA,…

Data disposal/retention policies must be systematically developed and enforced to benefit and protect organizations and individuals.

Before we continue, 4 important notes

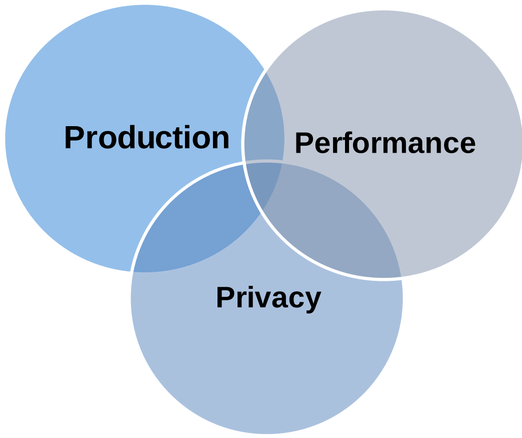


1) Not all data is important!



…

Before we continue, 4 important notes



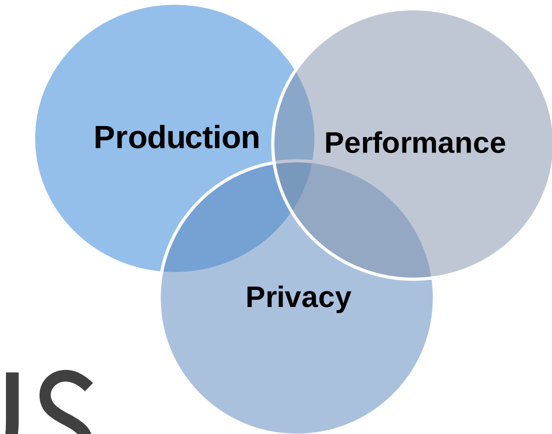
1) Not all data is important!

2) People fear of loosing potentially important data

3) Already now, sometimes there is really no choice

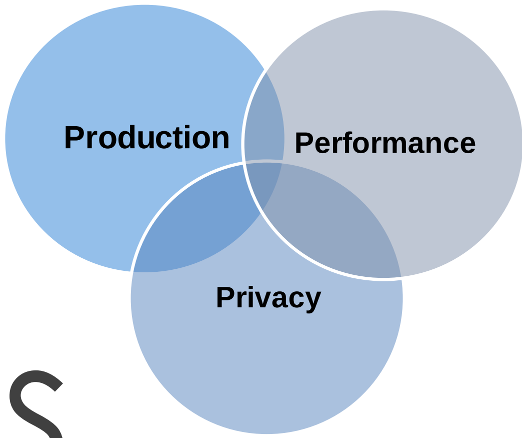
4) Like most good ideas, we are not the first to think about this …

Martin Kersten,
"The Wildest Idea" Award,
CIDR'15 Gong Show,
for "Big Data Space Fungus"

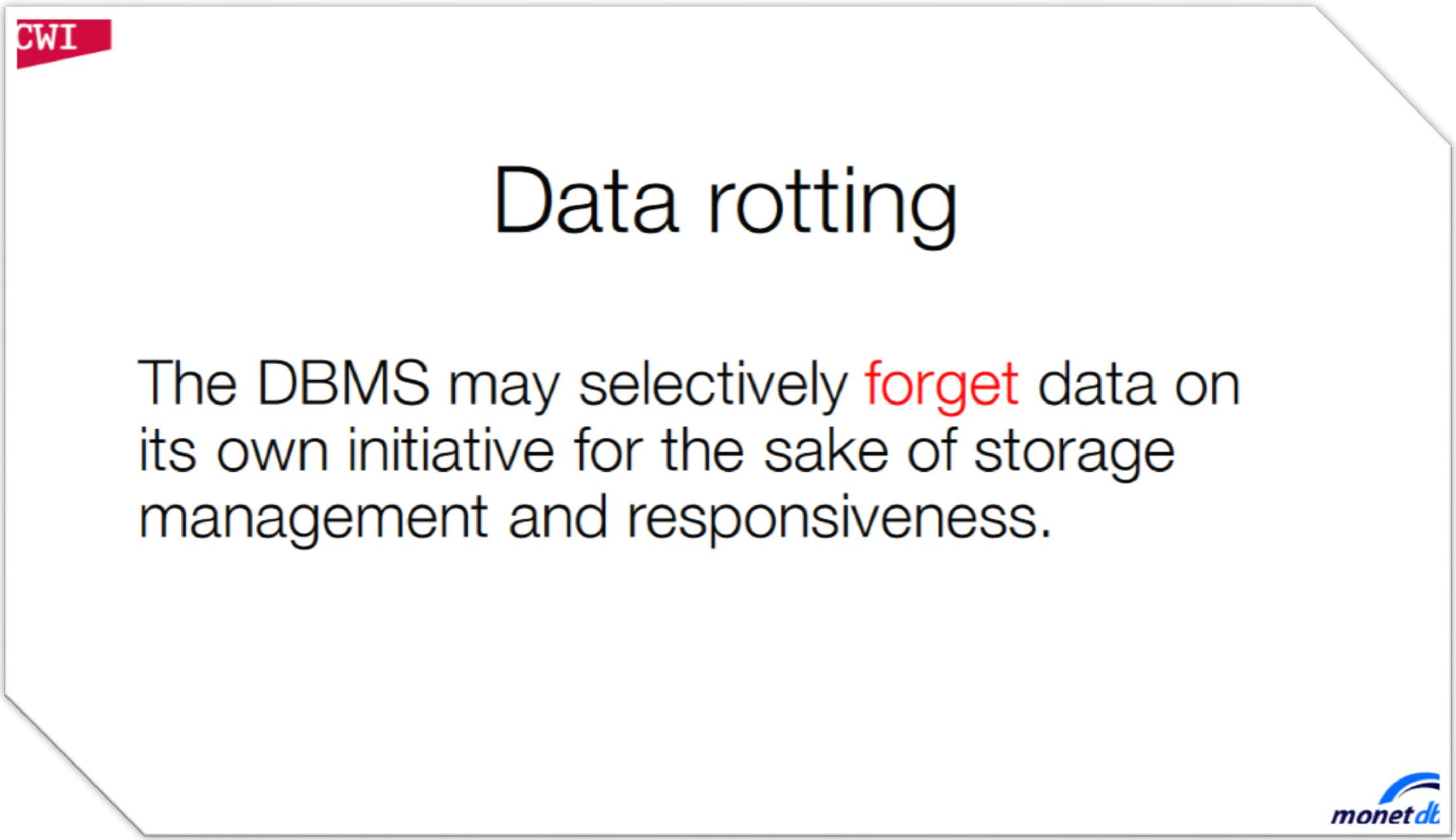


Big Data Space Fungus

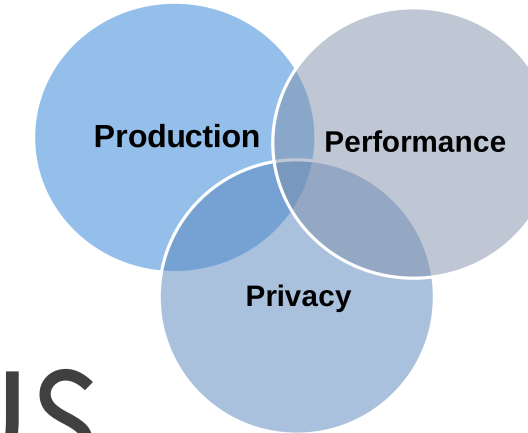
[CIDR'15]



Big Data Space Fungus



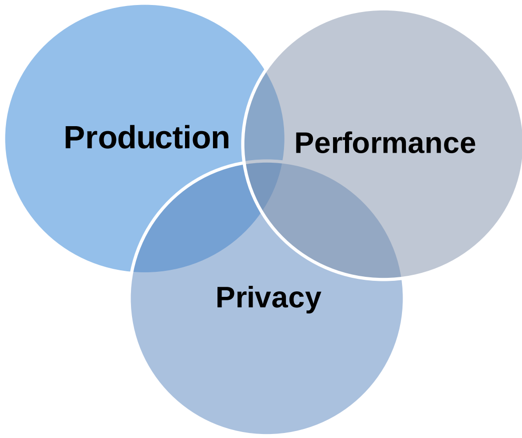
[CIDR’15]



Big Data Space Fungus

[CIDR'15]

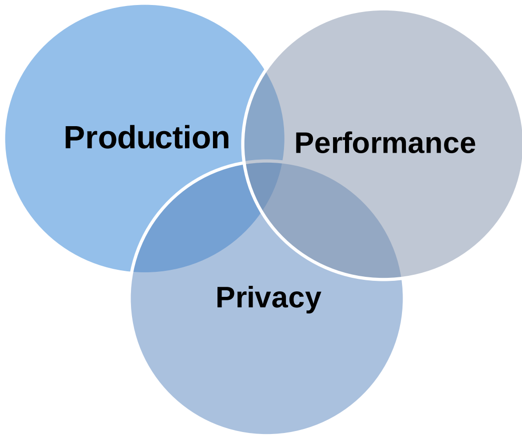
The Data Disposal Challenge



Retaining the knowledge hidden in the data while respecting storage, processing and regulatory constraints

- Determine an optimal **disposal policy** (which data to retain, summarize, dispose off) and execute it efficiently
- Support full-cycle information processing over the partial data
- Incrementally maintain the partial data as new info comes in

The 7 Criteria for Disposing Data



- What makes a piece of data important?
- How importance changes over time?
- Which of the data is important?
- Which data can (or must) be retained/disposed off? When?
- What is the cost of retaining / disposing off the data ?
- How can data be summarized / disposed off?
- How to process the partial data?

The Rest of This Talk

1. Existing tools
   (and why they are not enough)

2. Understanding the past
   (provenance)

3. Predicting the future
   (Deep Reinforcement Learning)

(Very) Incomplete List

Deduplication

- Entity resolution

(Semantic) compression & summarization

- Relations
- Semi-structured (XML, RDF, graph)
- Unstructured (text)

Sampling

- Approximate Query Processing

Sketching

- Streams

Machine Learning

- Dimensionality reduction
- Clustering
- Features selection

Example 1: Relations

Back to the late 90's…

age	salary	assets	credit	sex
20	30,000	25,000	poor	male
25	76,000	75,000	good	female
30	90,000	200,000	good	female
40	100,000	175,000	poor	male
50	110,000	250,000	good	female
60	50,000	150,000	good	male
70	35,000	125,000	poor	female
75	15,000	100,000	poor	male

(a) An Example Table

RRid	Bitmap	Outlying Values
2	01011	20, 25,000
1	11011	75,000
1	11111	
1	01100	40, poor, male
1	01111	50
1	01110	60, male
2	11110	female
2	11111	

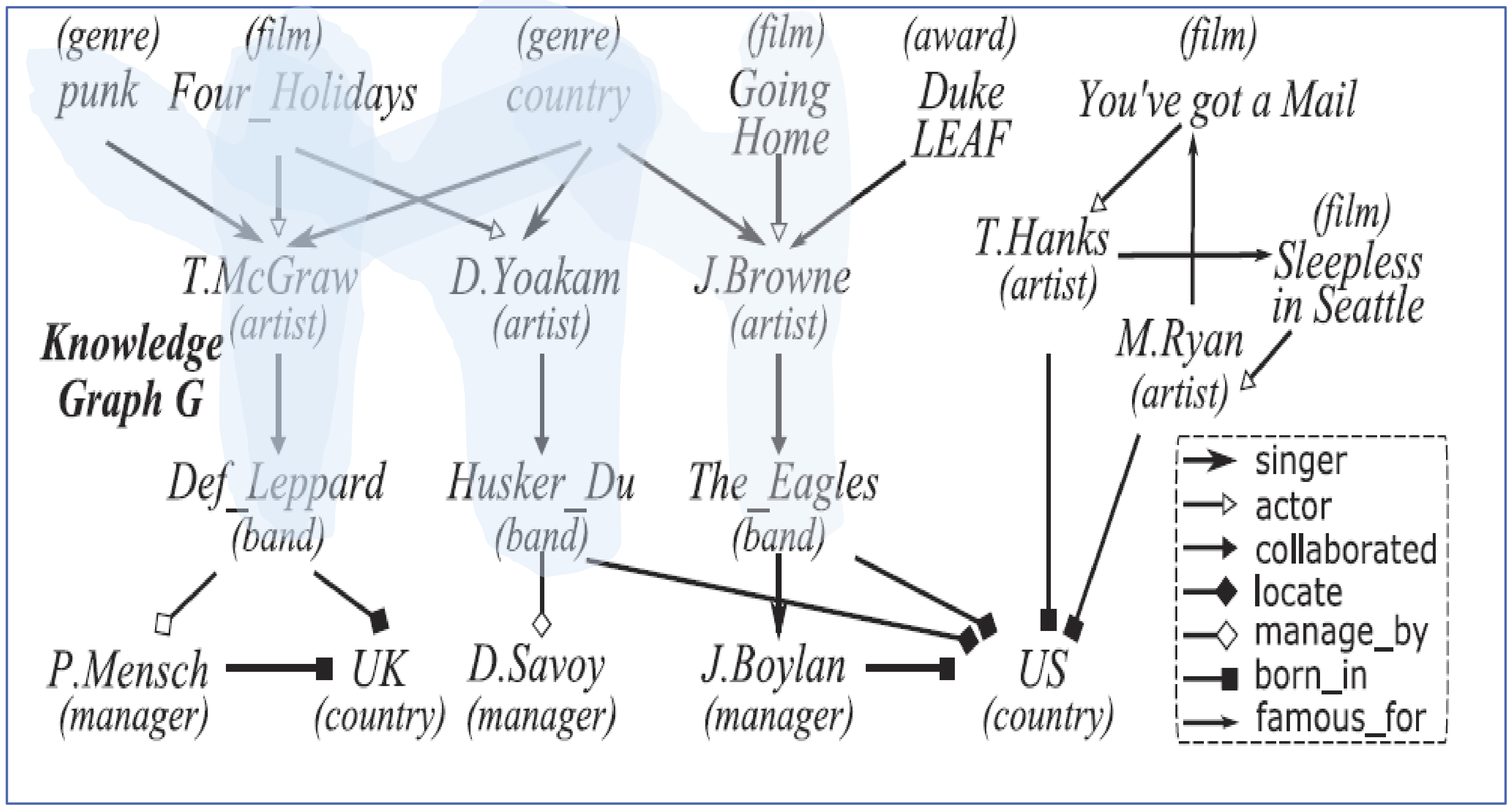
(b) Table T_c

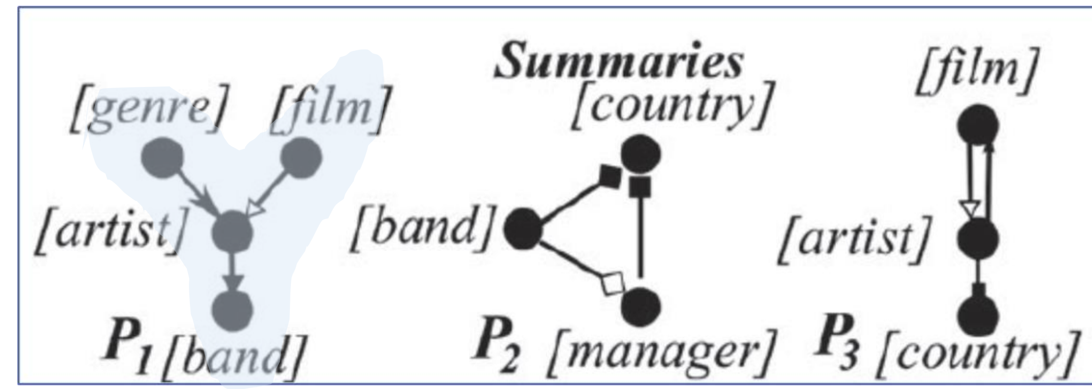
RRid	age	salary	assets	credit	sex
1	30	90,000	200,000	good	female
2	70	35,000	100,000	poor	male

(c) Representative Rows

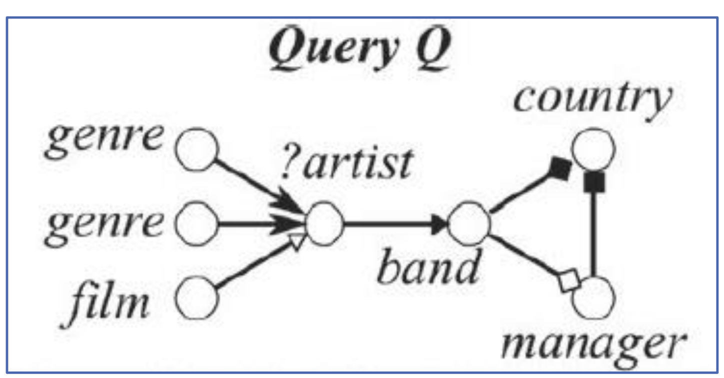
[Jagadish, Ng, Ooi, Tung, ICDE'04]

Example 2: Graphs





summary node	entities
[genre]	{ country, punk }
[film]	{Going Home, Four_Holidays}
[artist]	{J. Browne, D. Yoakam, T. McGraw}
[band]	{The_Eagles, Husker_Du, Def_Leppard }



[Song, Wu, Lin, Dong, Sun, TKDE'18]

Example 3: Sampling for AQP

Approximate query answers, at a fraction of full execution cost

- In query-time sampling, the query is evaluated over samples taken from the database at run time.
- For a sharper reduction on response time, draw samples from the data in a pre-processing step

  [Chaudhuri, Ding, Kandula, SIGMOD'17]

Question 1: Sample also from the data summaries?

Question 2: Use the precomputed samples as data summaries, thereby allowing to discard some (or all) of the remaining items?

Common Objectives

Summary properties

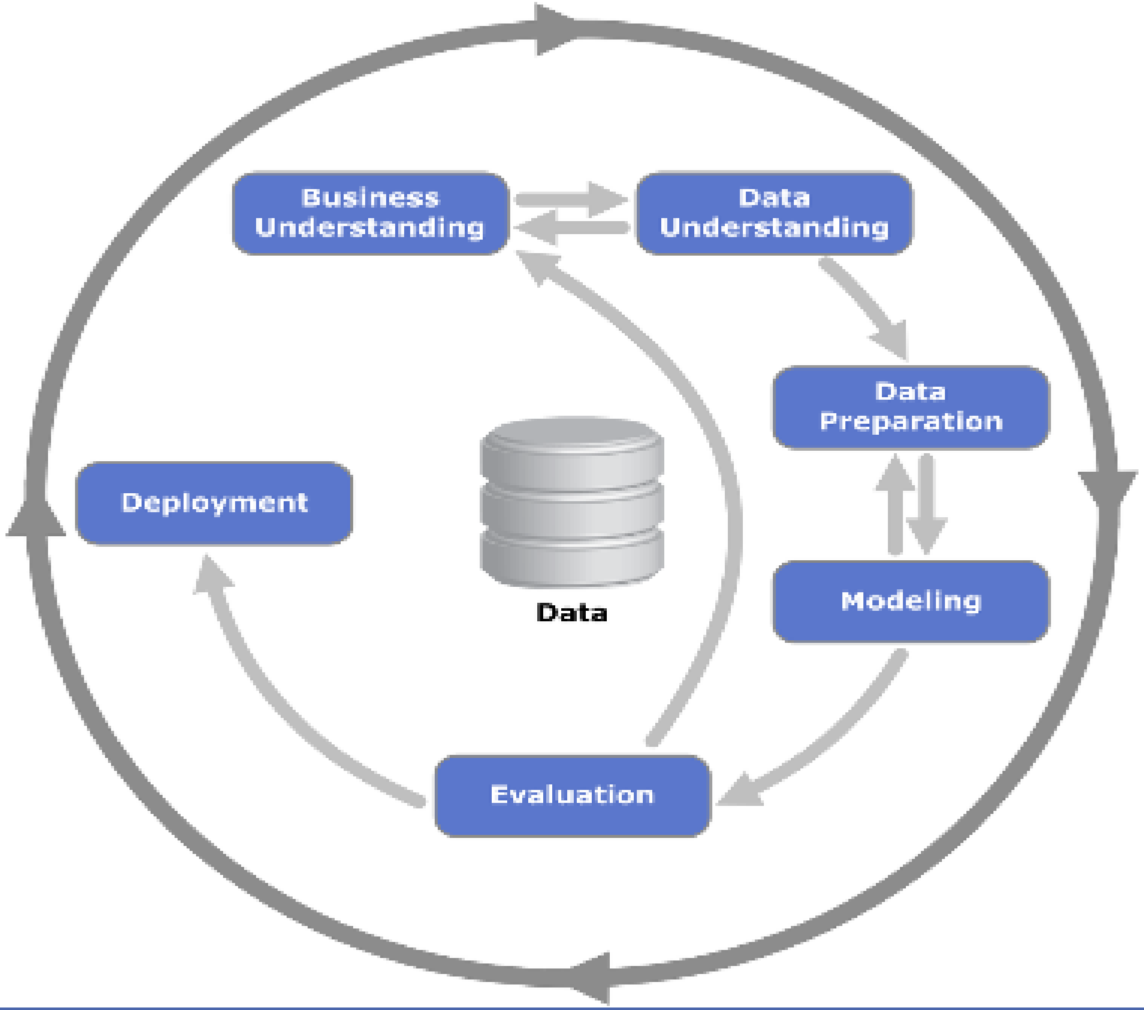
- Conciseness
- Diversification
- Coverage

Accuracy w.r.t **query results**

- Concrete queries
- Queries class/workload
- Information loss **[Orr, Suciu, Balazinska, VLDB'17]**

But in Practice…

Workloads are far more complex
(cleaning, transformation, integration, ML,…)



But in Practice…

Workloads are far more complex
(cleaning, transformation, integration, ML,…)

Need to understand how data is manipulated, summarized, disposed off throughout the entire workload !

The Rest of This Talk

1. Existing tools
   (and why they are not enough)

2. Understanding the past
   (provenance)

3. Predicting the future
   (Deep Reinforcement Learning)

Data Provenance

- Tracks computation and reveals the “origin” of results
- Many different models with different granularities
- Can be a key for performing & understanding data reduction

Provenance by Example

Customers

CID	Name	ZipCode
1	Lisa	99999
2	Homer	99999
3	Marge	99998
4	Bart	99999

CustLoans

CID	LID
1	1
1	2
1	3
2	4

Loans

LID	LoanType	Amount	Status	Date
1	UG Student Loan	50K	Denied	2017
2	Personal	100K	Denied	2017
3	Mortgage	85K	Approved	2018
4	G Student Loan	70K	Approved	2018

How many customers had a loan application denied in 2017 and accepted in 2018, per zip code?

```
SELECT C.ZipCode , COUNT(DISTINCT C.CID)
FROM Customers C, Loans L1, Loans L2, CustLoans CL
WHERE C.CID = CL.CID AND CL.LID = L1.LID AND CL.LID = L2.ID AND L1.Date = ‘2018’
            AND L2.Date = ‘2017’ AND L1.Status = ‘Approved’ AND L2.Status = ‘Denied’
GROUP BY C.ZipCode
```

Lineage

Customers

CID	Name	ZipCode
1	Lisa	99999
2	Homer	99999
3	Marge	99998
4	Bart	99999

CustLoans

CID	LID
1	1
1	2
1	3
2	4

Loans

LID	LoanType	Amount	Status	Date
1	UG Student Loan	50K	Denied	2017
2	Personal	100K	Denied	2017
3	Mortgage	85K	Approved	2018
4	G Student Loan	70K	Approved	2018

How many customers had a loan application denied in 2017 and accepted in 2018, per zip code?

Lineage tells us that Marge's and Bart's info does not contribute to the analysis output, and hence may be, or must be (by GDPR!) removed

Provenance Polynomials

Customers

CID	Name	ZipCode
1	Lisa	99999
2	Homer	99999
3	Marge	99998
4	Bart	99999

CustLoans

CID	LID
1	1
1	2
2	4
1	3

Loans

LID	LoanType	Amount	Status	Date
1	UG Student Loan	50K	Denied	2017
2	Personal	100K	Denied	2017
3	Mortgage	85K	Approved	2018
4	G Student Loan	70K	Approved	2018

How many customers had a loan application denied in 2017 and accepted in 2018, per zip code?

The provenance Polynomial include, for 99999:

....+ Customers(1,Lisa,99999) * [CustLoans(1,1) * Loans(1,UG,50K,Denied, 2017)
+ CustLoans(1,2) * Loans(2,Morgage,100K,Denied, 2017)]
* CustLoans(1,3)
* Loans(3,Personal,80K,Approved,2018) + ...

Provenance Polynomials

Customers

CID	Name	ZipCode
1	Lisa	99999
2	Homer	99999
3	Marge	99998
4	Bart	99999

CustLoans

CID	LID
1	1
1	2
2	4
1	3

Loans

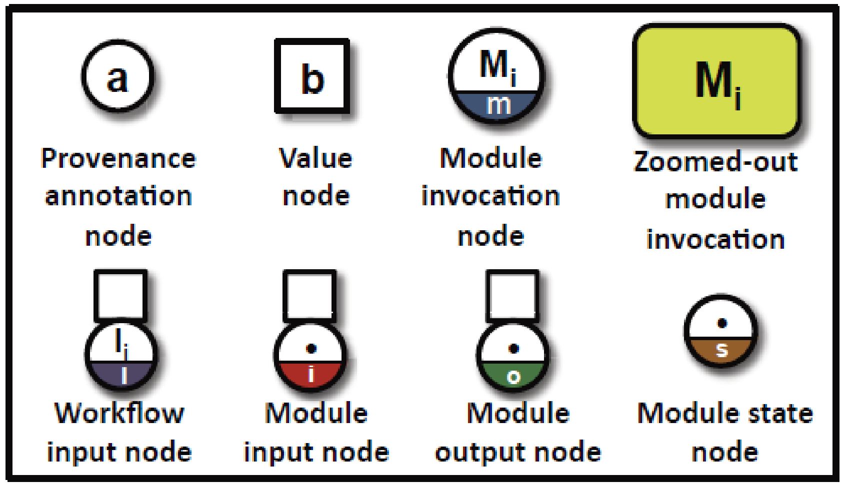
LID	LoanType	Amount	Status	Date
1	UG Student Loan	50K	Denied	2017
2	Personal	100K	Denied	2017
3	Mortgage	85K	Approved	2018
4	G Student Loan	70K	Approved	2018

One of these may also be deleted

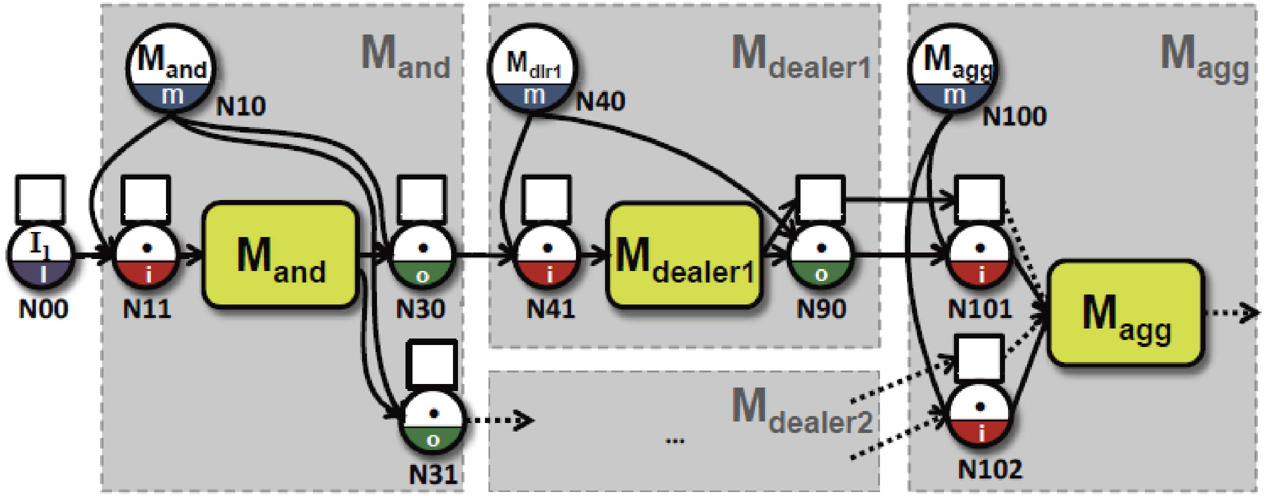
The provenance Polynomial include, for 99999:

....+ Customers(1,Lisa,99999) * [CustLoans(1,1) * Loans(1,UG,50K,Denied, 2017)
+ CustLoans(1,2) * Loans(2,Morgage,100K,Denied, 2017)]
* CustLoans(1,3)
* Loans(3,Personal,80K,Approved,2018) + ...

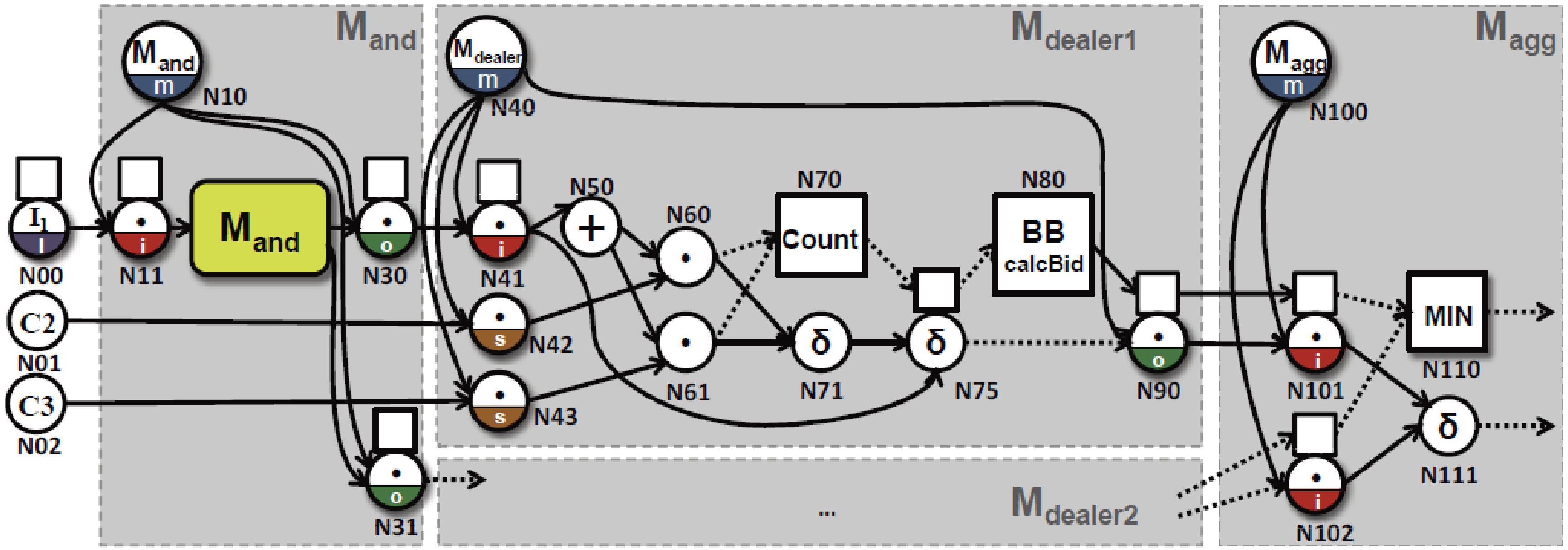
Workflow Provenance



(a) Legend



(b) Coarse-grained provenance



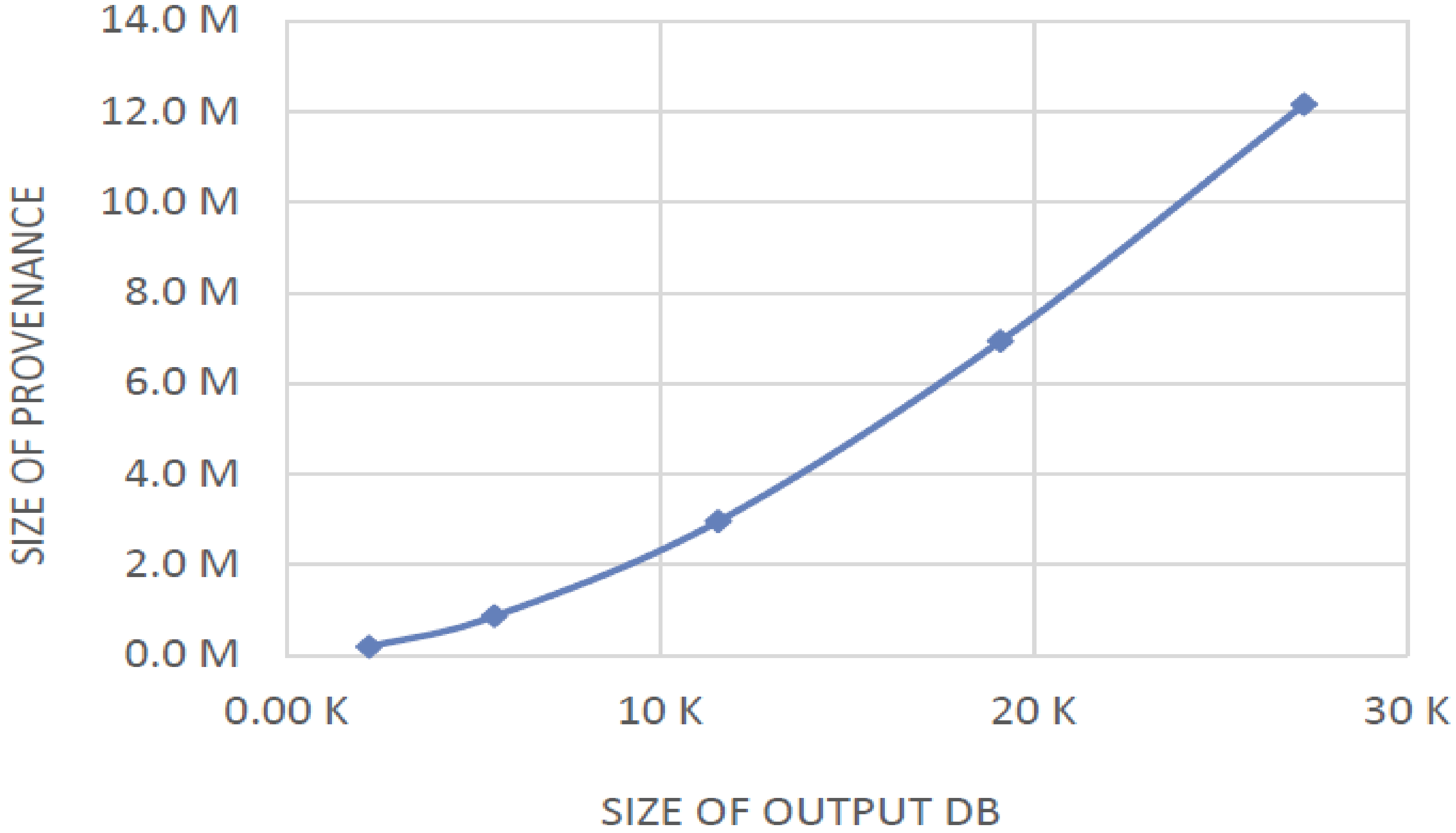
(c) Fine-grained provenance

Many Applications

- Results Explanation
- Hypothetical reasoning
- Trust level assessment
- Computation in presence of incomplete/probabilistic info.
- Data reduction **[Gershtein, M, Novgorodov, CIKM'19]**
- …

But...

Provenance is **HUGE**



Datalog provenance, for a 3-rule recursive program

Provenance Reduction

Lossless

- Size reduction via expression simplification/factorization (e.g. using Boolean circuits)

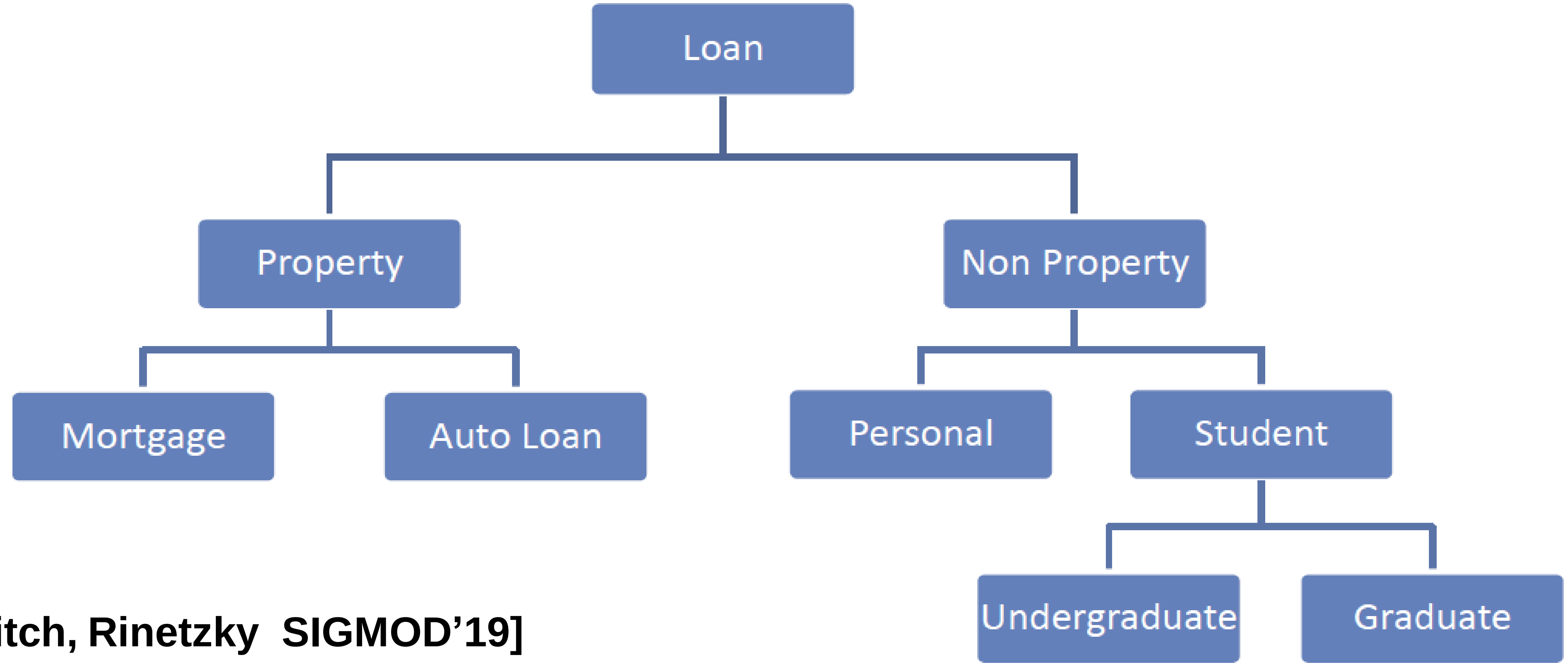
Lossy

- Selective provenance
- Compression via abstraction

Example: Compression by Abstraction

How many customers had a loan application denied in 2017 and accepted in 2018, per zip code?

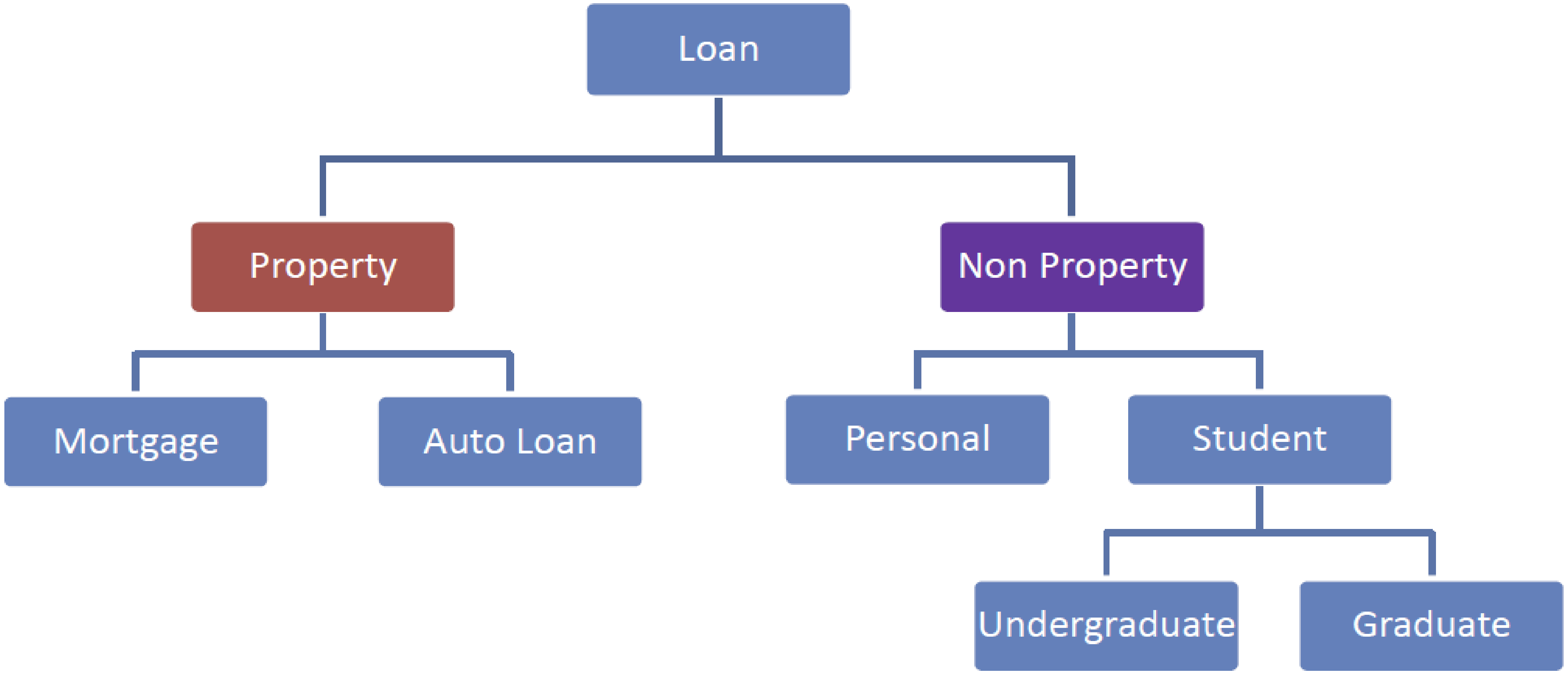
Maybe we do not need to store individual loan requests, but just an abstraction?



[Deutch, Moskovitch, Rinetzky SIGMOD'19]

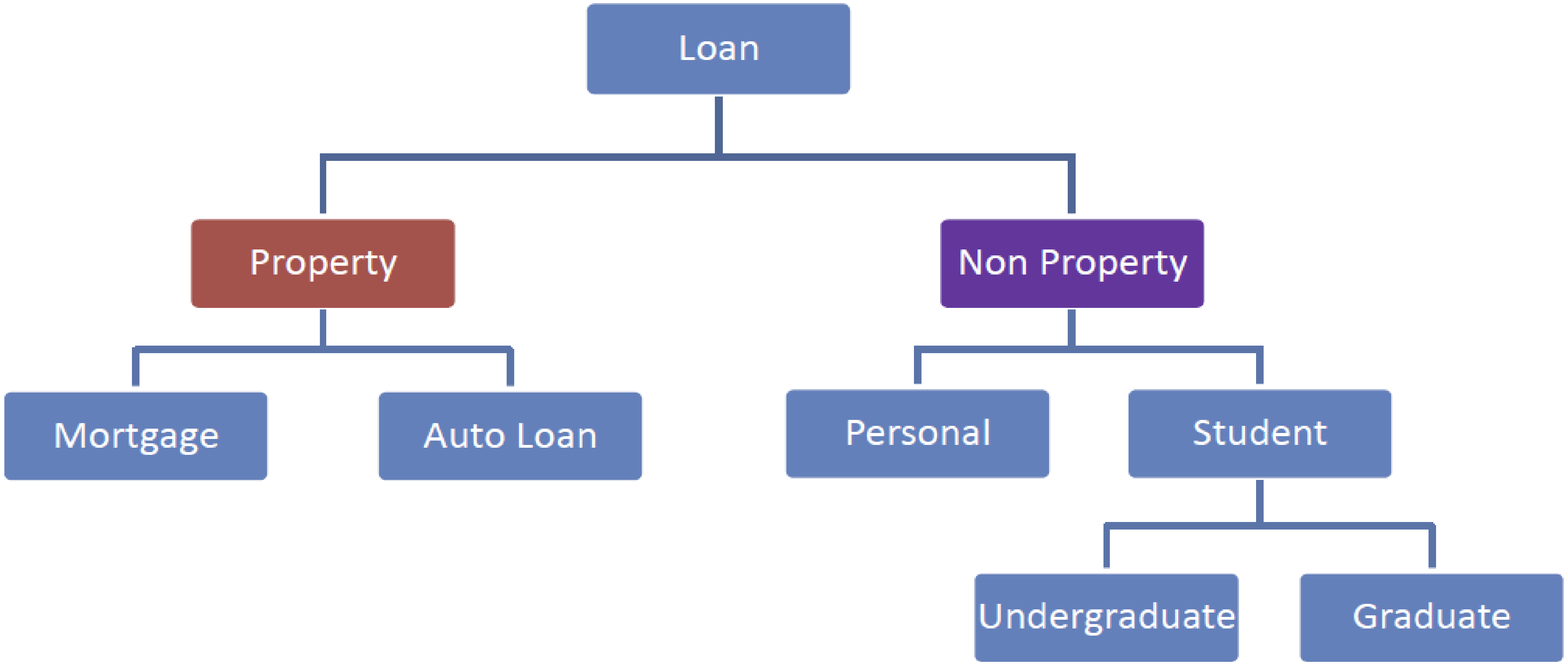
Example: Compression by Abstraction

....+ Customers(1,Lisa,99999)* [CustLoans(1,1) ·Loans(1,UG,50K,Denied, 2017)
+ CustLoans(1,2) ·Loans(2,Personal,100K,Denied, 2017)]
* CustLoans(1,3) * Loans(3,Mortgage,80K,Approved,2018)



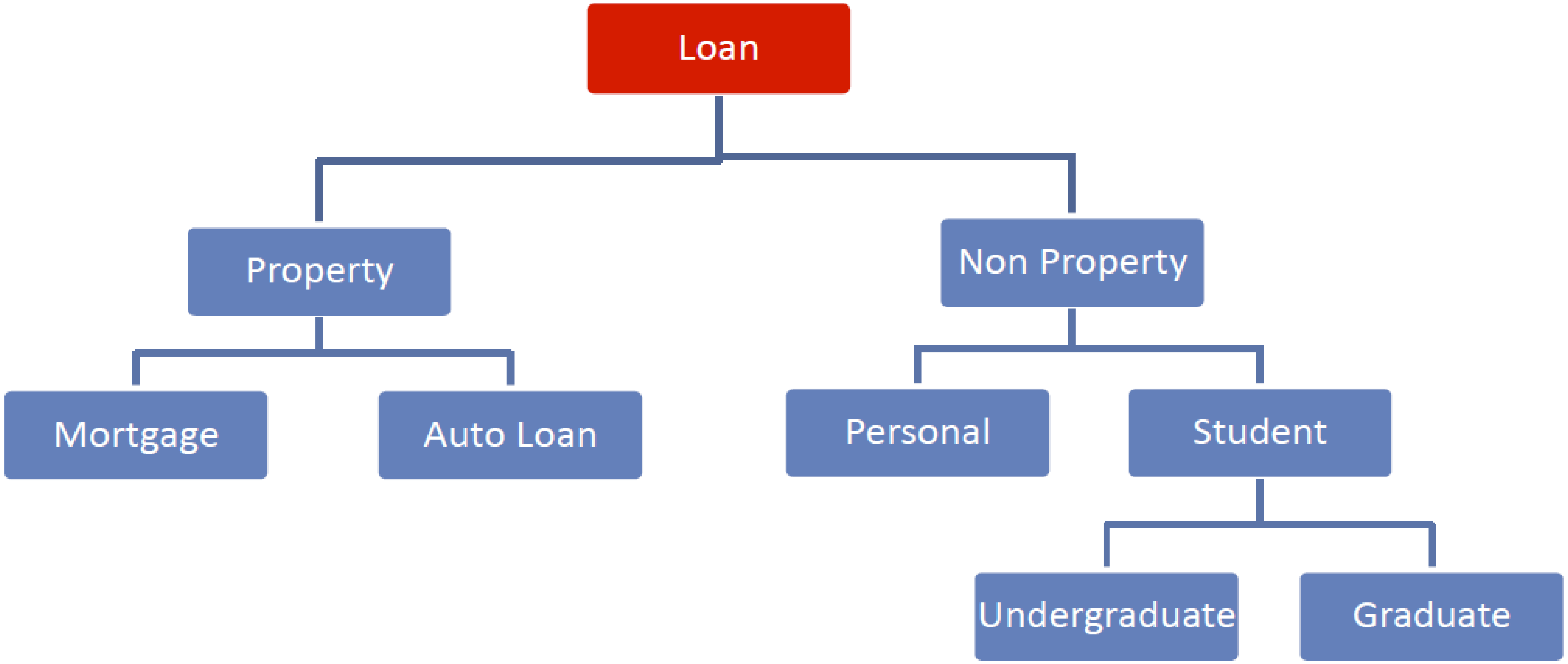
Example: Compression by Abstraction

....+ Customers(1,Lisa,99999)* [2 NonProperty]
* Property

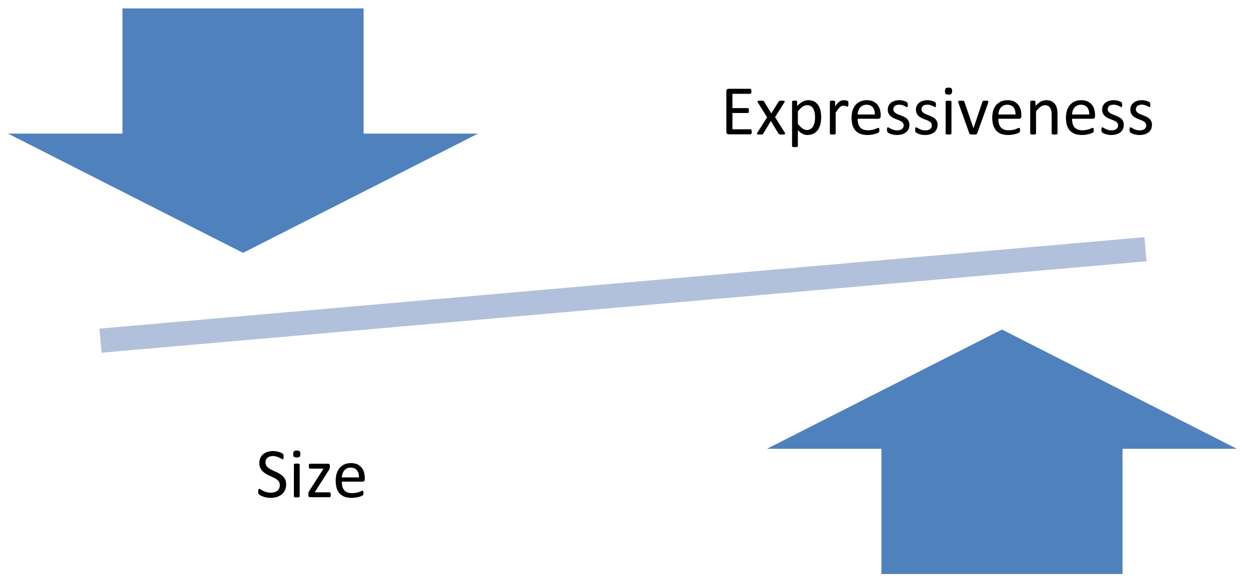


Example: Compression by Abstraction

....+ Customers(1,Lisa,99999) * **2 Loan ^2**



Optimization Problem



- Choose a cut in the ontology that maximizes expressiveness for a target compression ratio
- **NP-hard** in general
- **Polynomial** time complexity for a **single ontology**
- Practically appealing heuristics for the general case

The Rest of This Talk

1. Existing tools
   (and why they are not enough)

2. Understanding the past
   (provenance)

3. Predicting the future
   (Deep Reinforcement Learning)

Learn what may be interesting in a new dataset

Exploratory data analysis (EDA):

The process of examining & investigating a given dataset

Exploratory Data Analysis

EEDA is an *iterative* process:

- A **user u** loads a dataset **D** to an analysis interface.
- Performs a sequence of: $S_U(D) = q_1, q_2, \ldots q_n$ of actions (e.g. queries)
- After executing q_i - the user examines the results, and decides if and which action to perform next.

The goal:

- Understand the nature of the dataset
- Discover its properties
- Estimate its quality
- **Figure our what may be interesting in it**

Modern analysis platforms (e.g. Splunk, Kibana-ELK, Tableau, …)

EDA agent

Can we teach a machine to generate a coherent, meaningful sequence of exploratory queries?

Deep Reinforcement Learning

DRL works surprisingly well for very difficult tasks:

- Play Go
- Drive a car
- Conduct natural language dialogs
- ……

Can/Should we use DRL?

PROS:

- It requires NO training data OR traces of user activity
- Once trained - results can be obtained rather FAST.

CONS:

- It is a heavy-weight tool, requires lots of computing power.
- Currently works mostly on game-like environments
- Even when working - it may just overfit to some odd patterns in the data

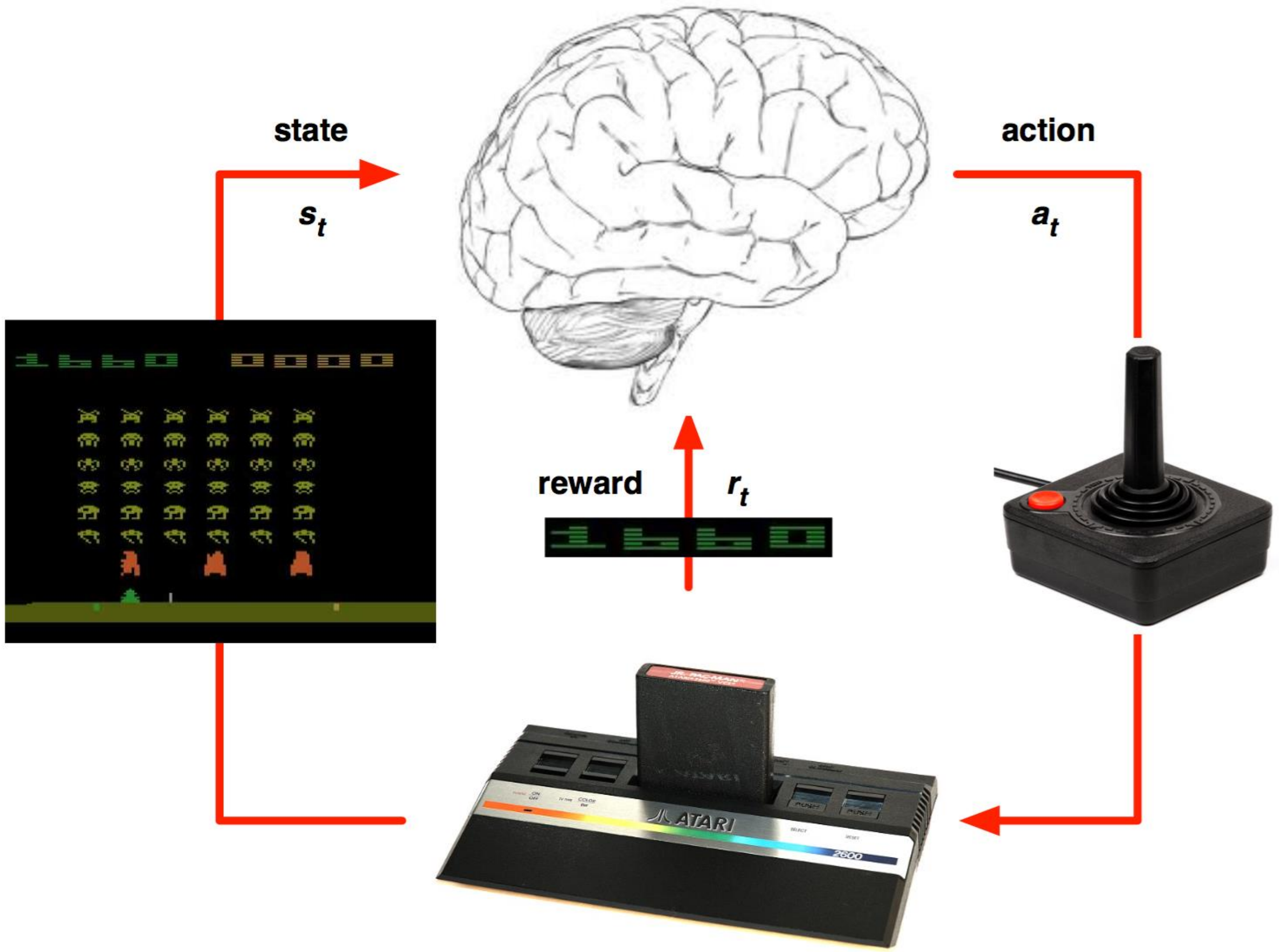
The Rest of This Talk

1. Quick recap of standard RL settings
2. Requirements for RL-EDA environment
3. Our framework (ongoing work)

RL Standard Settings

In the (not so simple) Atari environment:

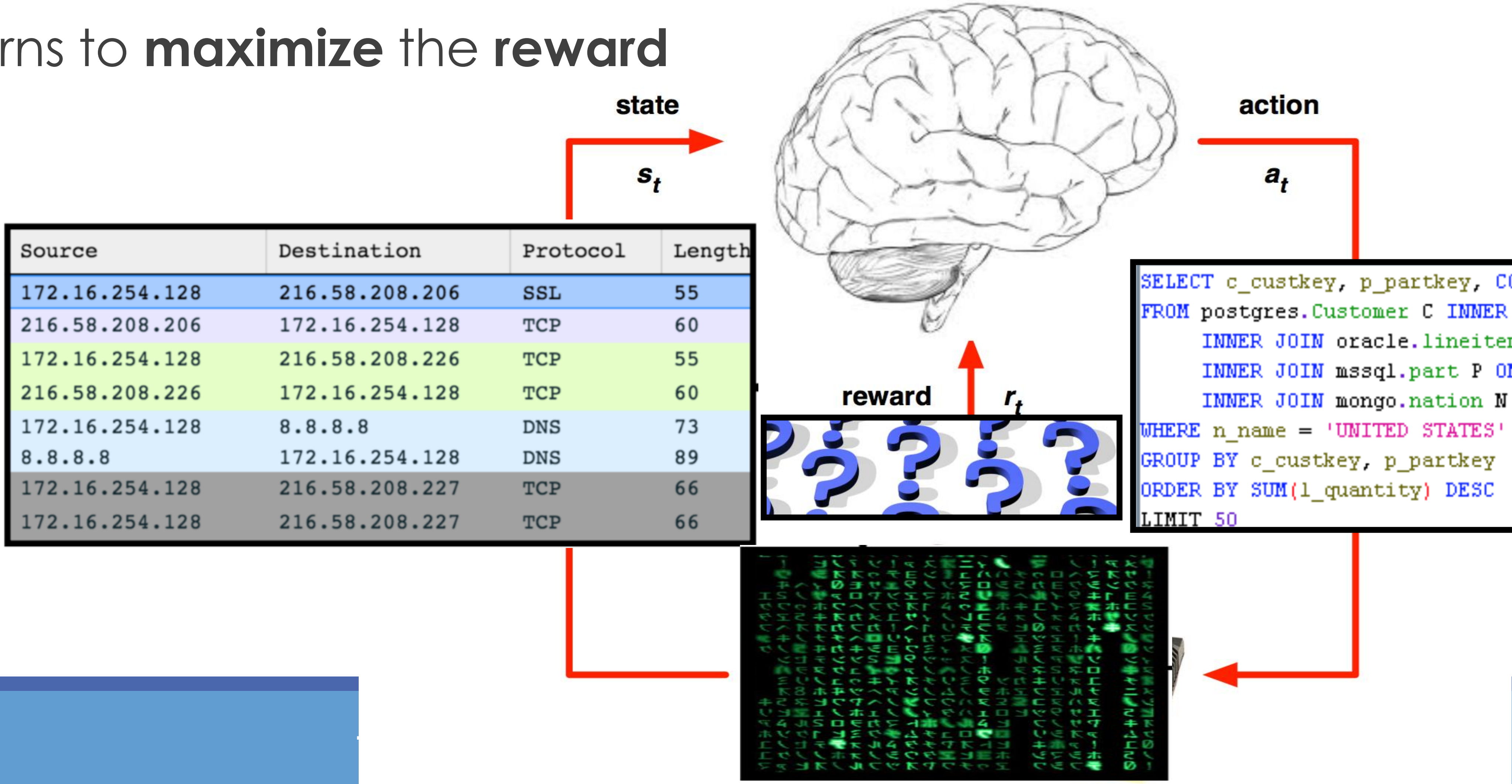
1. Agent observes a “**State**” from an “**environment**”
2. Agent selects an “**action**”
3. Agent receives “**reward**”
4. Agent learns (unsupervised) a “**policy”** that **maximizes** the **mean reward**



RL-EDA Settings

Utilizing the RL paradigm for EDA:

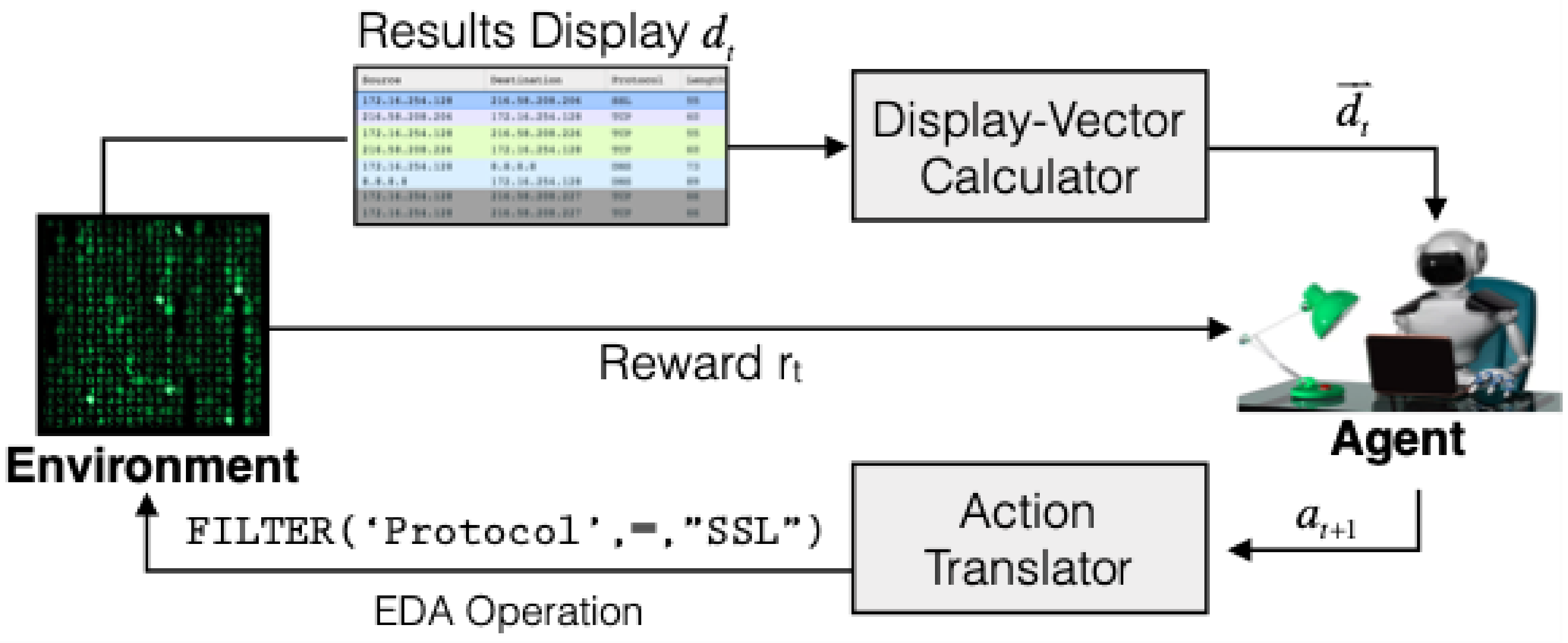
1. Agent observes a **dataset/results set**
2. Agent formulates a **query**
3. Agent receives **reward**
4. Agent learns to **maximize** the **reward**



Source	Destination	Protocol	Length
172.16.254.128	216.58.208.206	SSL	55
216.58.208.206	172.16.254.128	TCP	60
172.16.254.128	216.58.208.226	TCP	55
216.58.208.226	172.16.254.128	TCP	60
172.16.254.128	8.8.8.8	DNS	73
8.8.8.8	172.16.254.128	DNS	89
172.16.254.128	216.58.208.227	TCP	66
172.16.254.128	216.58.208.227	TCP	66

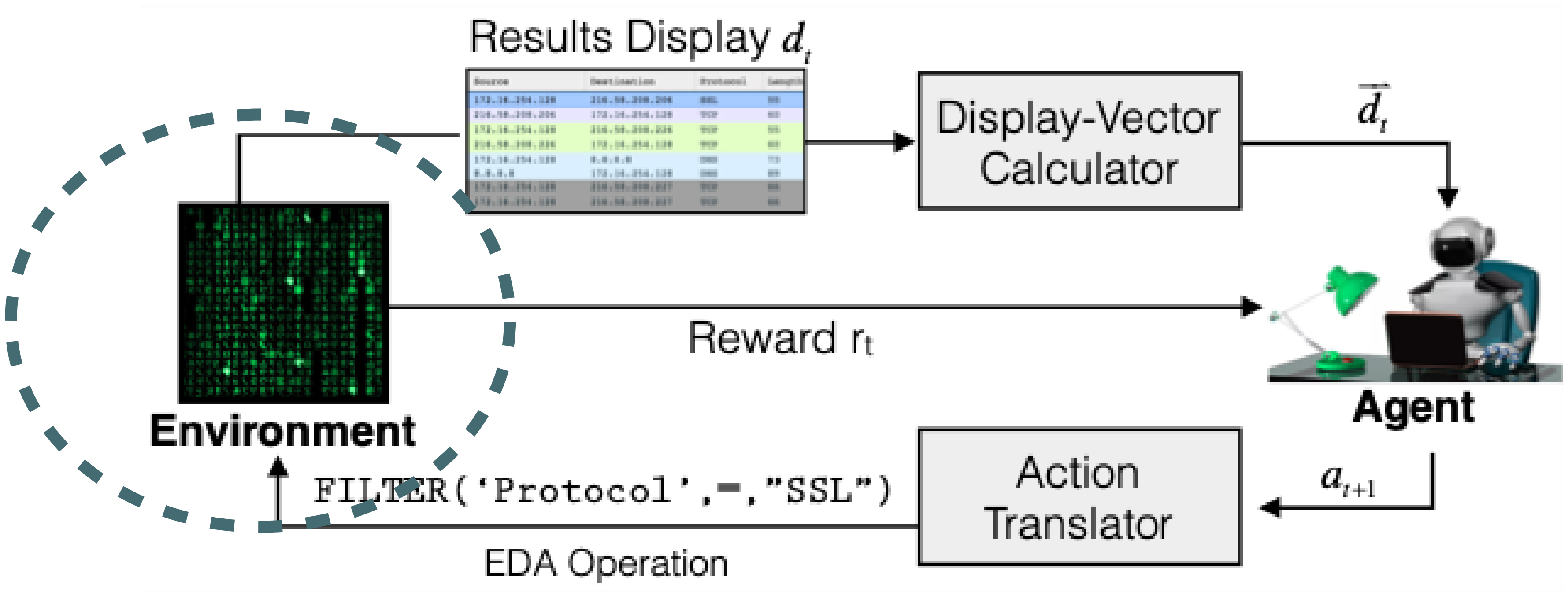
Outline for an RL-EDA Framework

1. RL-EDA environment
2. State and action representation
3. Reward Signal
4. Agent NN-Architecture



Outline for an RL-EDA Framework

1. **RL-EDA environment**
2. State and action representation
3. Reward Signal
4. Agent NN-Architecture



RL-EDA Environment

RL-EDA environment comprises:

(1) A collection of datasets

(2) Query interface

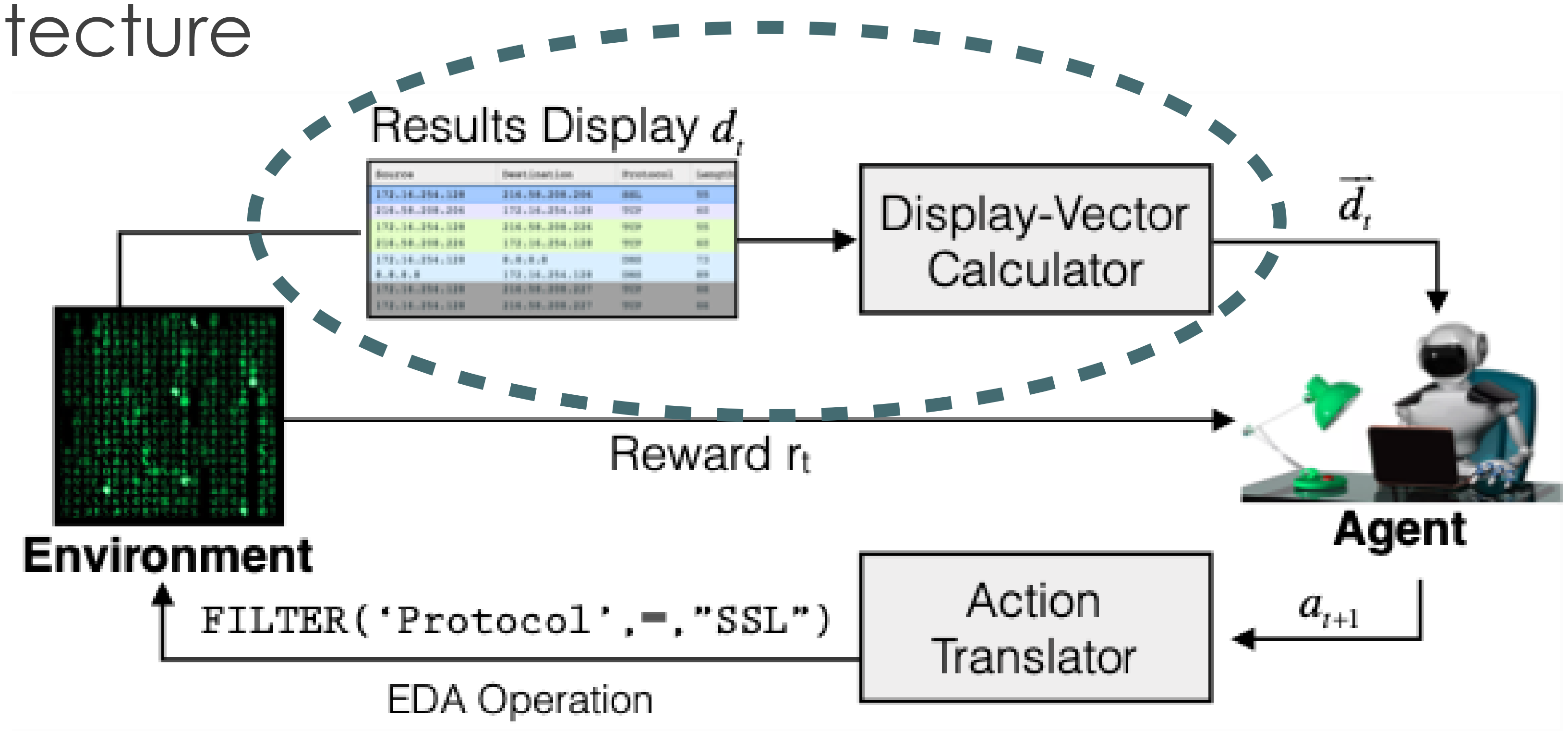
RL-EDA Episode:

The agent is "given" an arbitrary dataset

The agent performs a "session" (sequence) of N queries.

Outline for an RL-EDA Framework

1. RL-EDA environment
2. **State** and action **representation**
3. Reward Signal
4. Agent NN-Architecture



State Representation

Result displays are often large and complex…

→ Summarize the results display into a numeric vector

- <u>Structural features of the data:</u>

  Value entropy, # of distinct values, # of Null values

- <u>Grouping/Aggregation features:</u>

  # of groups, groups size variance, aggr. values, entropy,…

- <u>Context:</u>

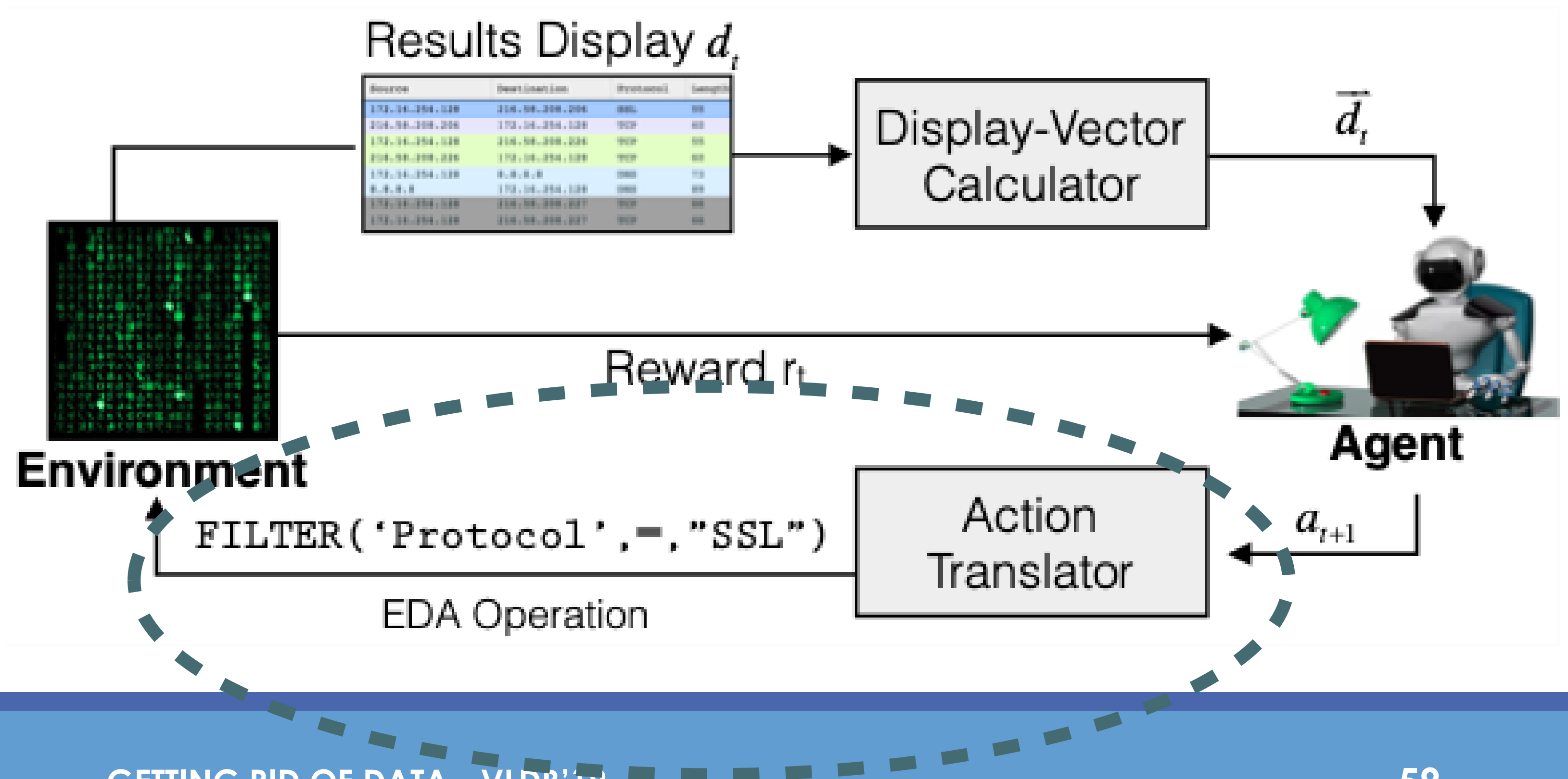
  N previous displays

Source I P	Pa..	Lenth	Time Stamp	Source M A C	Dest M A C	Dest I P	Sour..	Dest ..	T C..	Protoc..
										SMB
192.150.11.111 (5)		**avg: 187.2**								
192.150.11.111	13	143	1240208908	00:30:48:62:4e:4a	00:08:e2:3b:56:01	98.114.205.102	445	1828	1	SMB
192.150.11.111	16	311	1240208909	00:30:48:62:4e:4a	00:08:e2:3b:56:01	98.114.205.102	445	1828	1	SMB
192.150.11.111	19	175	1240208909	00:30:48:62:4e:4a	00:08:e2:3b:56:01	98.114.205.102	445	1828	1	SMB
192.150.11.111	22	114	1240208909	00:30:48:62:4e:4a	00:08:e2:3b:56:01	98.114.205.102	445	1828	1	SMB
192.150.11.111	25	193	1240208909	00:30:48:62:4e:4a	00:08:e2:3b:56:01	98.114.205.102	445	1828	1	SMB
98.114.205.102 (5)		**avg: 199.8**								
98.114.205.102	10	191	1240208908	00:08:e2:3b:56:01	00:30:48:62:4e:4a	192.150.11.111	1828	445	1	SMB
98.114.205.102	14	222	1240208908	00:08:e2:3b:56:01	00:30:48:62:4e:4a	192.150.11.111	1828	445	1	SMB
98.114.205.102	17	276	1240208909	00:08:e2:3b:56:01	00:30:48:62:4e:4a	192.150.11.111	1828	445	1	SMB
98.114.205.102	20	152	1240208909	00:08:e2:3b:56:01	00:30:48:62:4e:4a	192.150.11.111	1828	445	1	SMB
98.114.205.102	23	158	1240208909	00:08:e2:3b:56:01	00:30:48:62:4e:4a	192.150.11.111	1828	445	1	SMB

$$\longrightarrow \begin{bmatrix} 0.96 \\ 0.32 \\ 0.91 \\ \cdots \\ 1 \\ 0 \\ 2 \\ 1 \\ \cdots \end{bmatrix}$$

Outline for an RL-EDA Framework

1. RL-EDA environment
2. State and **action representation**
3. Reward Signal
4. Agent NN-Architecture



Action Representation

Parameterized Actions (action type + parameters)

- FILTER(attr, op, term) - used to select data tuples that matches a criteria
- GROUP(attr, agg func, agg attr) - groups and aggregates the data
- BACK() - allows the agent to backtrack to a previous display

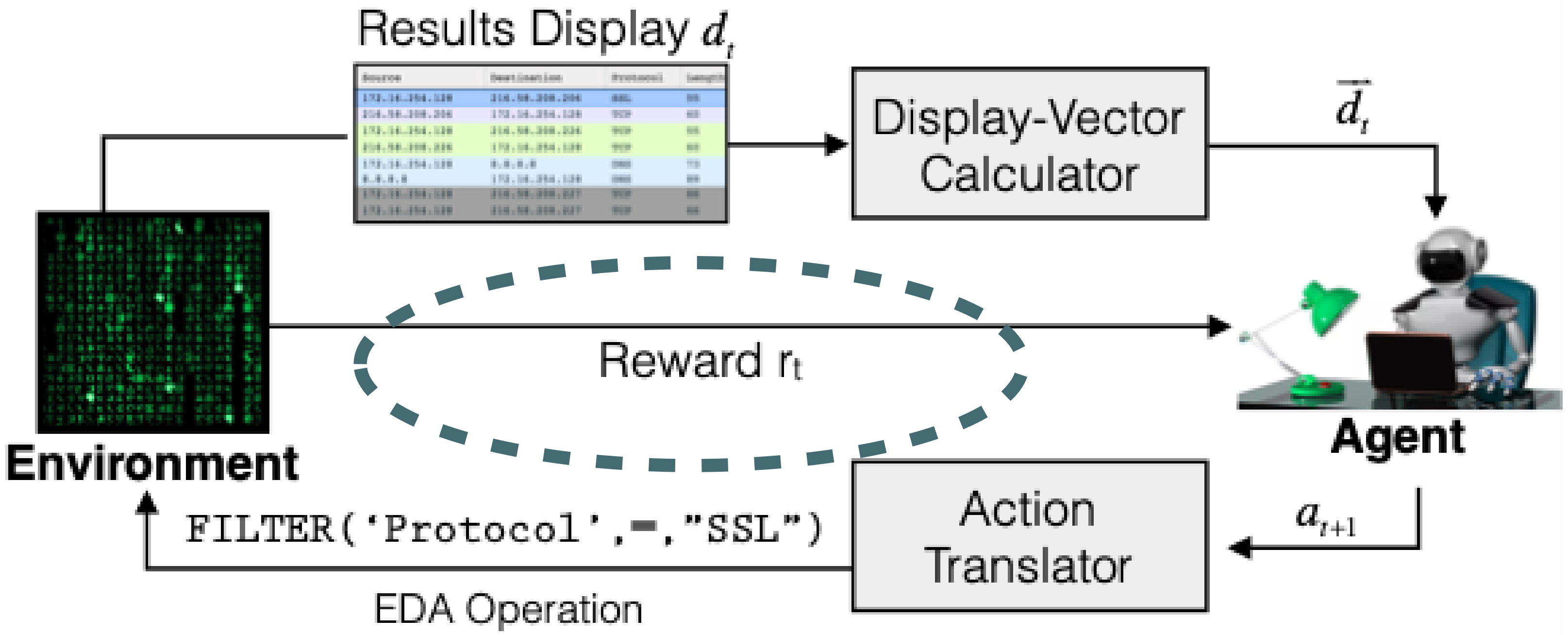
Our Representation

- [action_type, attr, op, term, agg_func, agg_attr]
- Handle filter terms using the frequency of appearances in the display

Issue: large actions domain

Outline for an RL-EDA Framework

1. RL-EDA environment
2. State and action representation
3. **Reward Signal**
4. Agent NN-Architecture



Reward Signal

Given a sequence S_D= q_1, q_2,…q_n of queries performed by the agent on dataset D. How to determine the reward $R(S_D)$?

We suggest three major components.

1. **Interestingness:** Actions inducing **interesting** results set should be encouraged
2. **Diversity:** Actions in the same session should yield **diverse** results describing different aspects of the dataset
3. **Coherency:** The session is understandable to human analysts

Interestingness

Multitude of interestingness measures are suggested in previous work.

Each captures a different aspect of interestingness:

Diversity	Measures how much the elements of a data pattern are different from on another
Pecularity	Measures how *anomalous* is a pattern comparing to the rest of the data patterns
Conciseness	Measures the size of the pattern compared to its coverage
Novelty	Measures how *unexpected* a data pattern is w.r.t. known prior knowledge

Diversity

Goal: encourage the agent to choose actions inducing new observations of different parts of the data than those examined so far

Solution: calculate the Euclidean distances between the observation vector of the current results display and the vectors of all previous displays

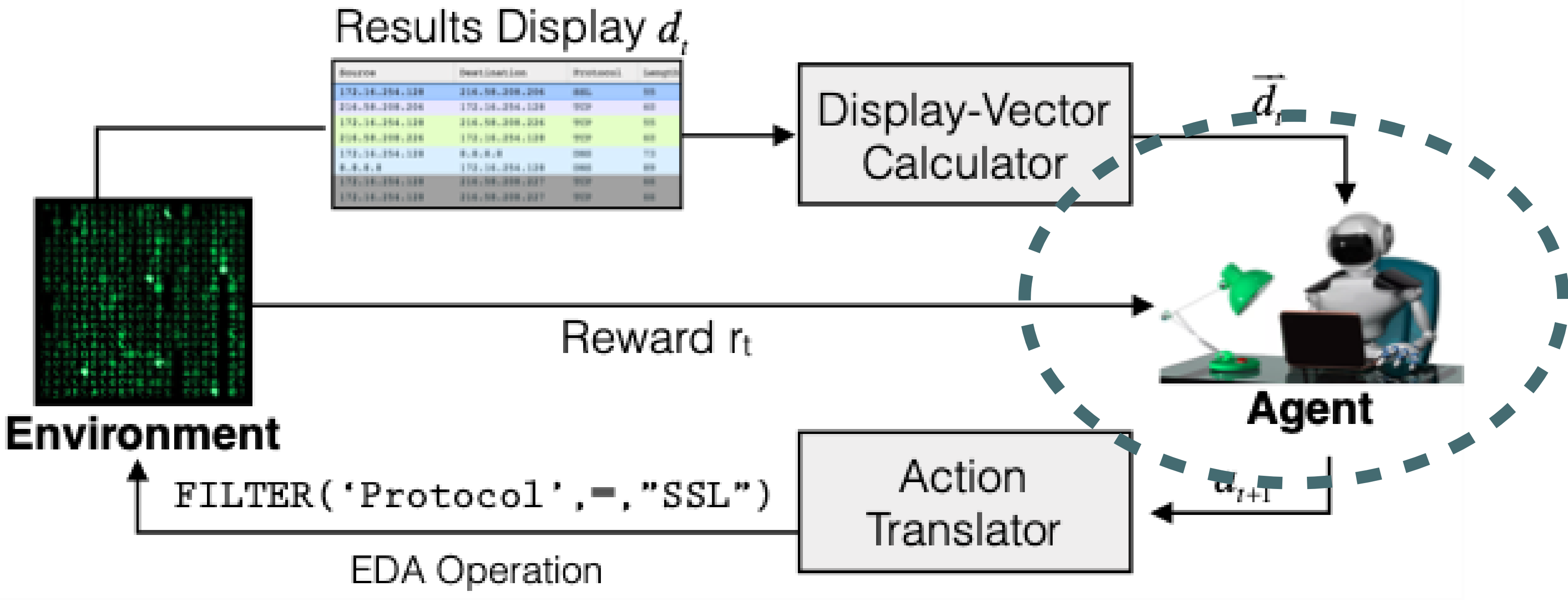
Coherency

Performed using an external classifier:

1. Given the dataset schema & application domain we use a set of **heuristic classification-rules composed by domain experts** (e.g. “a group-by that is employed on more than 4 attributes is non-coherent”)
2. Then employ **Snorkel** to build a weak-supervision based classifier

Outline for an RL-EDA Framework

1. RL-EDA environment
2. State and action representation
3. Reward Signal
4. **Agent NN-Architecture**

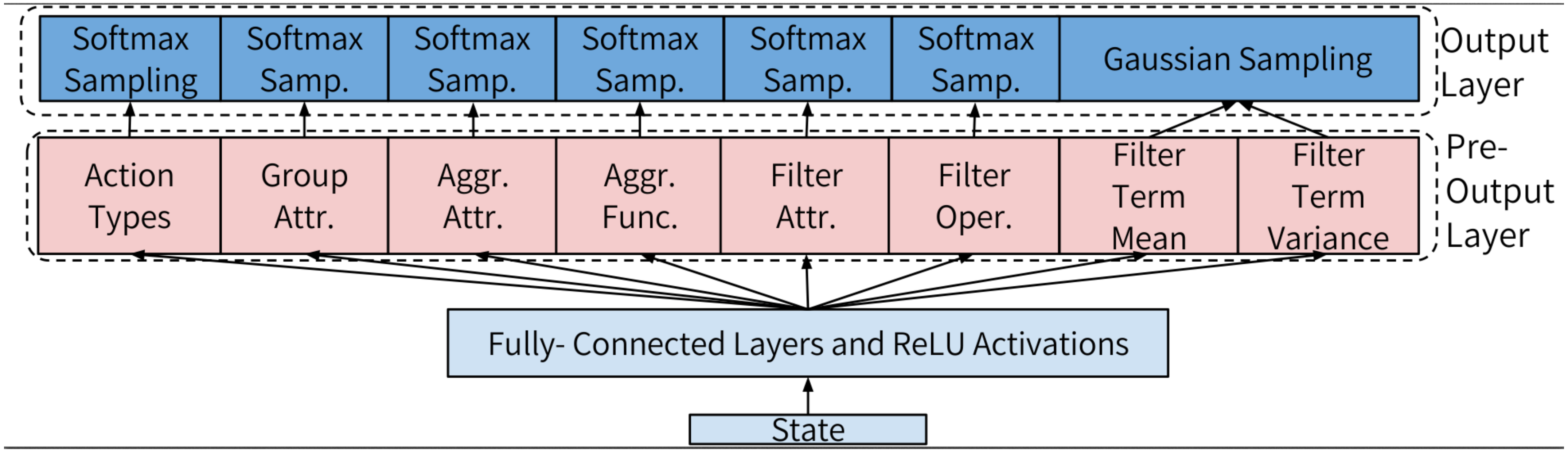


Challenges

Large # of actions
(in particular due to the Filter parameter)

Exploration challenges: imbalanced action types
(BACK, GROUP, FILTER)

Our solution: parameterized softmax with pre-output layer



A few words about experimental evaluation

1. Learning curves and reward

2. Competitors: Greedy, Recommender systems, Human…

3. Measures: BLEU, sessions similarity
"Turing test"

Time to Conclude…

Time to Conclude…

The Data Disposal Challenge

- Determine an optimal **disposal policy** (which data to retain, summarize, dispose off) and execute it efficiently

- Support full-cycle information processing over the partial data

- Incrementally maintain the partial data as new info comes in

Define formally what makes a disposal policy good…

Time to Conclude…

1. Plenty of relevant tools

2. But still **very** far from a comprehensive solution

3. ML agents: Still a lot to do here!

    - Support more data analysis actions
    - Adaptive disposal policies based on user interaction
    - Consider potential data exploration goals

Thank You

Ori Bar-El, Naama Boer, Daniel Deutch, Shay Gershtein, Amir Gilad, Gefen Keinan, Nave Frost, Yuval Moskovitch, Slava Novgorodov, Kathy Razmadze, Noam Rinetzky, Amit Somech, Brit Youngmann, …