Towards an Introspective Dynamic Model of Globally Distributed Computing Infrastructures

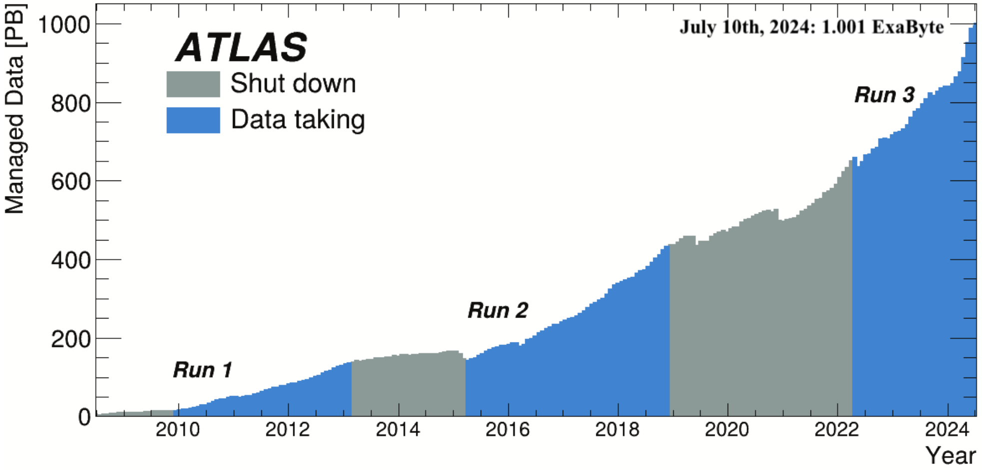
Yihui Ren[1], Ozgur Kilic[1], David Park[1], Tatiana Korchuganova[6], Frederic Suter[2], Joseph Boudreau[6], Norbert Podhorszki[2], Paul Nilsson[1], Sairam Sri Vatsavai[1], Scott Klasky, Shengyu Feng[5], Tasnuva Chowdhury[7], Varena Ingrid Martinez Outschoorn[4], Yiming Yang[5] , Tadashi Maeno[1], Alexei Klimentov[1], Adolfy Hoisie[1]

[1]Brookhaven National Laboratory, Upton, NY; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]SLAC National Accelerator Laboratory, Menlo Park, CA; [4]University of Massachusetts at Amherst, Amherst, MA; [5]Carnegie Mellon University, Pittsburgh, PA; [6]University of Pittsburgh, Pittsburgh, PA

CHEP 2024 Track 7 | Presenter: David Park

Project Goals and Organization

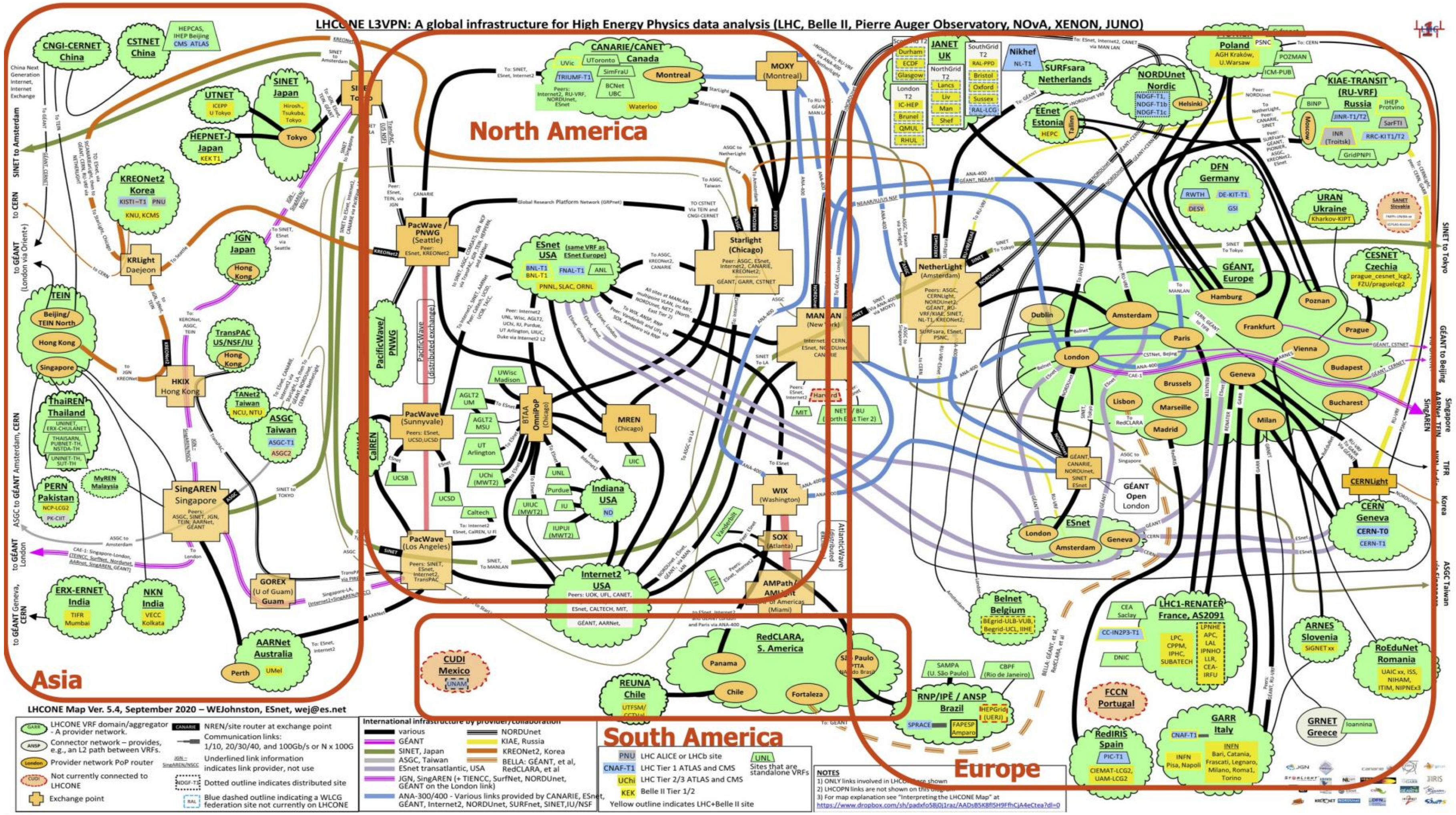
Distributed computing sites of global scientific collaborations



- Motivation: Extreme large data volumes and increasingly complex computation workflows in many scientific domains
- Goal: Optimal data placement and workload scheduling enhancing the resilience, throughput, and resource utilization.

World Nuclear and Particle Physics Research Network

WAN connectivity increased x10 in 10 years. This shows a Virtual Private Network (LHCONE) spanning150 sites in ~40 countries on all continents but Antarctica, and its bandwidth is dedicated to High Energy Physics.



Brookhaven National Laboratory

Four Interacting Components of the Dynamic Model

Real-time Feedback

Dispatcher: e.g. WMS PanDA breaks task into jobs and assign them to compute sites.

INPUT
(Surrogate model of INPUT)

Dispatcher (WMS PanDA)
(Break Task into Jobs, assign Jobs to sites)

OUTCOME
(reliability, queuing time, error rate, data movement)

INPUT: user and centrally managed tasks submitted to the distributed system.

OUTCOME: observables of the distributed system. We can use these observables to define the performance and reliability metrics, calibrate dynamic models and provide real-time feedback to the dispatcher.

Environment
(Sites, storage capacity, computing capacity, bandwidth, local queues)

Build an interactive dynamic model to represent the system and evaluate different policies.

Environment: different sites have different computing capabilities, storage, data partition, network bandwidth, local jobs, and failure rate, etc.

Brookhaven National Laboratory

Production and Distributed Analysis (PanDA)

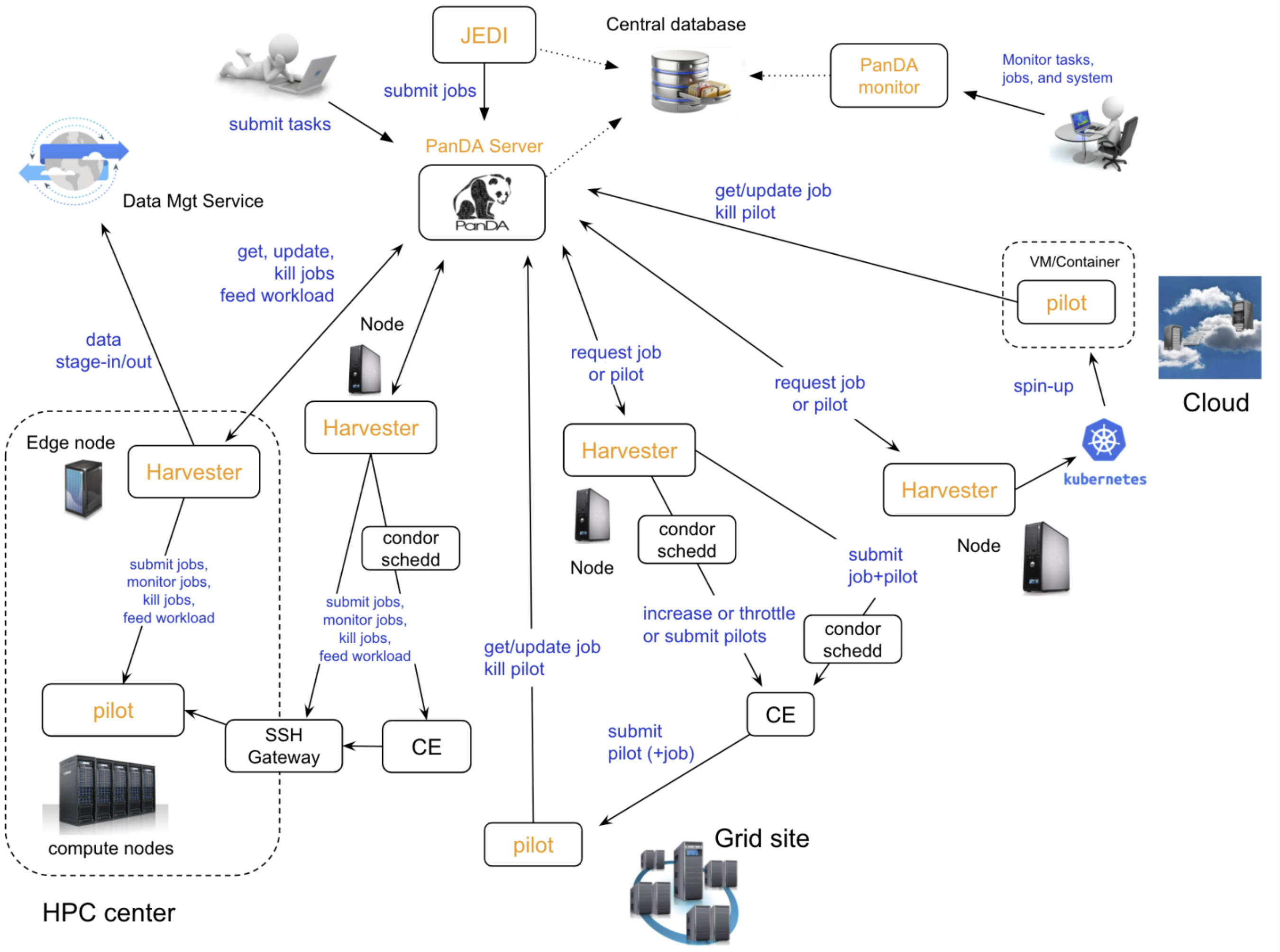
- The PanDA system has been developed by ATLAS since the summer of 2005 to address the experiment's need for a data-driven workload management solution capable of handling both production and distributed analysis at the scale required for LHC data processing.

- **Workflow**: a group of tasks; **Task**: a group of jobs

- A **job** runs on a slot in computing resource to process a subset of input and produce a subset of output.

- **Note**: "task" in some other systems means "job" in our terminologies

[PanDA] T. Maeno et al., "PanDA: Production and Distributed Analysis System," Comput. Softw. Big Sci., vol. 8, no. 1, p. 4, 2024.



PanDA system overview

Production and Distributed Analysis (PanDA)

PanDA

Dataset statistics

Time span: 150 days

(Jan 1, 2024 – June 1, 2024)

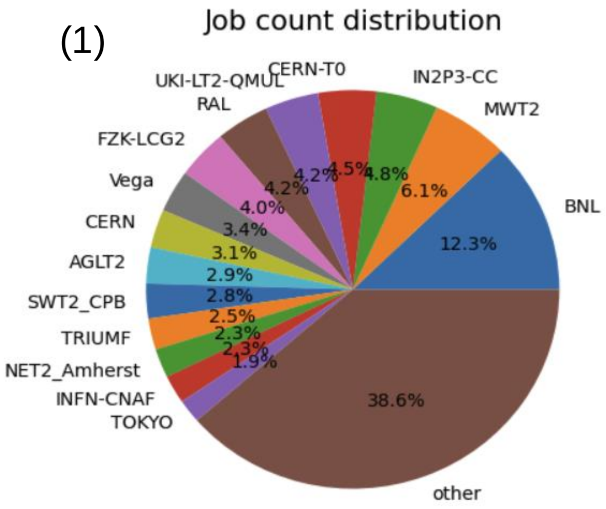
Number of user jobs: 2,352,392

Number of unique columns: 131

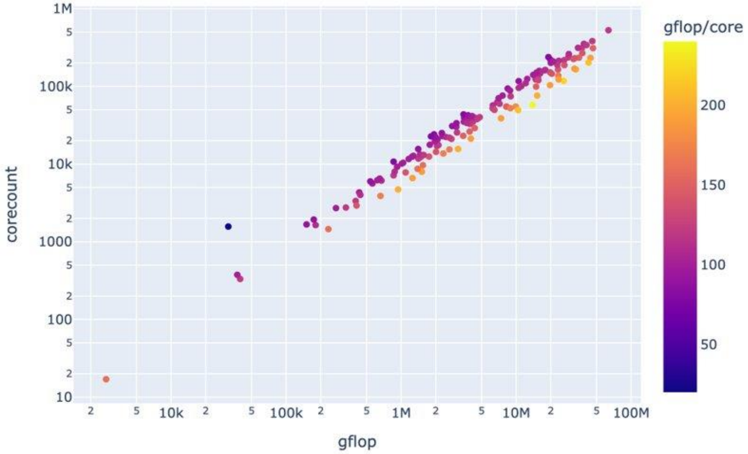
Number of unique tasks: 10990

- (Fig. 1) User jobs are distributed in multiple computing sites
- (Fig. 2) Computing sites show varying sums of FLOPs
- (Fig. 3) Most jobs finish successfully while some others fail.
- (Fig. 4) Median job takes 3100 CPU seconds.

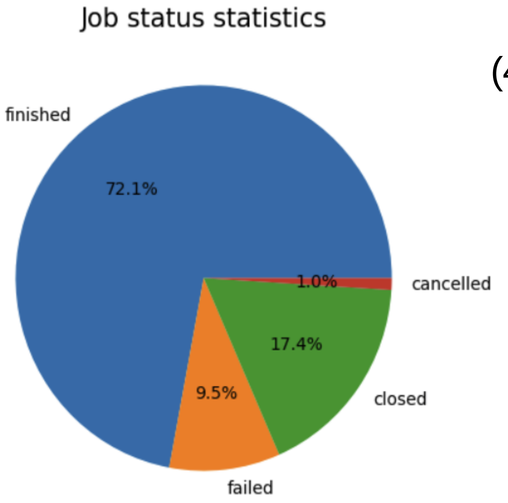
(1)



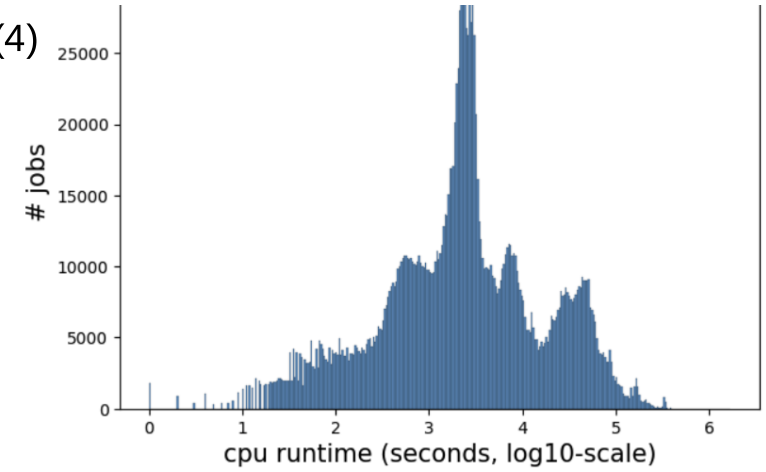
(2)



(3)



(4)



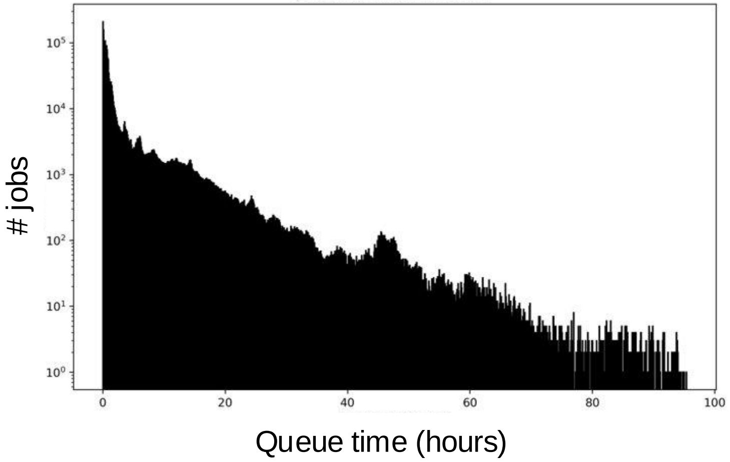
Brookhaven National Laboratory

Identification of key introspective metrics

Identified several introspective measures for resiliency

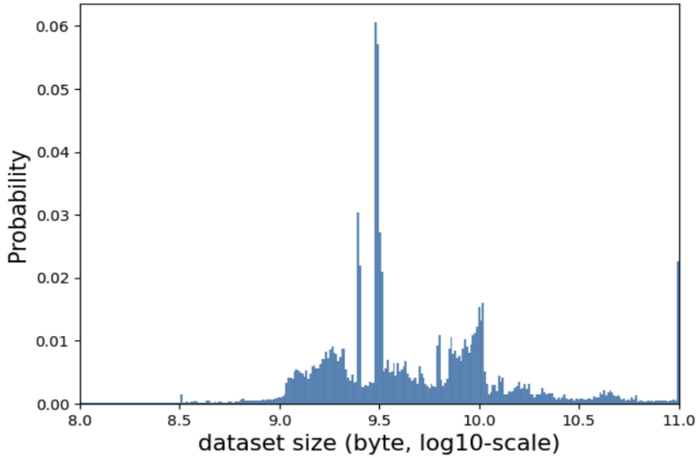
- Job queue time
- Wasted time due to errors
- Dataset sizes and movement

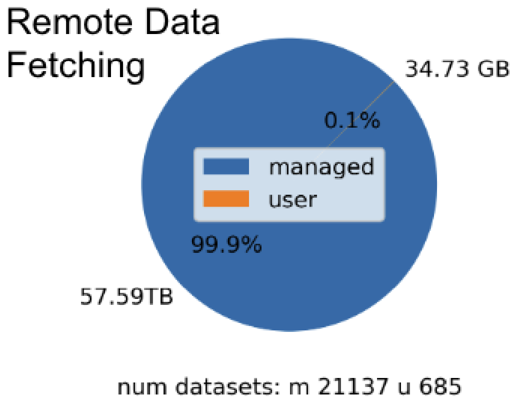
Job queue time distribution



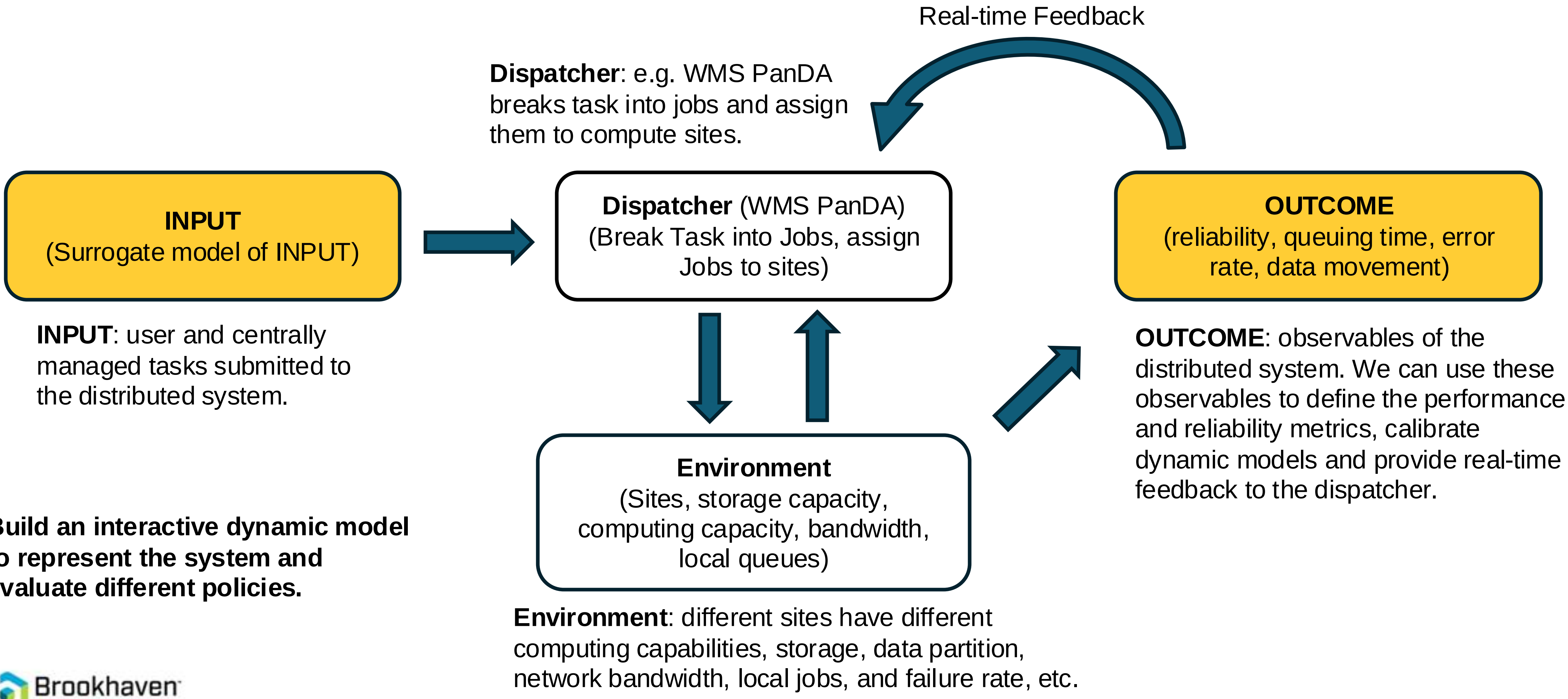
Wasted time per error

Density of dataset sizes





Four Interacting Components of the Dynamic Model



Build an interactive dynamic model to represent the system and evaluate different policies.

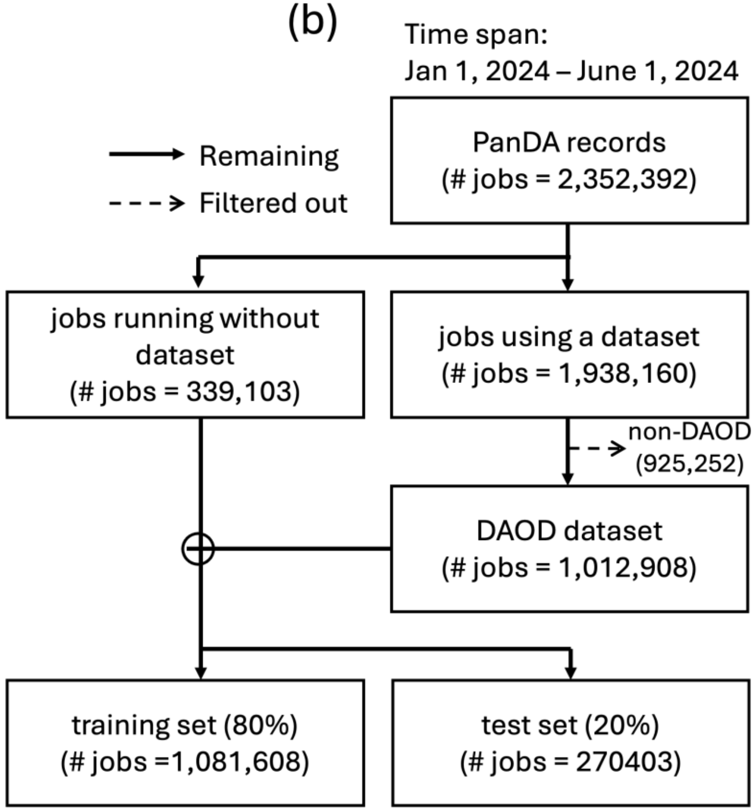
Representative features for surrogate modeling [1]

- Preprocessing pipeline (b) and preprocessed data samples (a).

(a)

	creation time	computing site	DAOD dataset features					status	workload
			project	prod step	data type	nfiles	size		
type	N	C	C	C	C	N	N	C	N
# unique	N/A	83	14	4	54	N/A	N/A	4	N/A
samples	2024-03-24 21:09:26	ANALY_BNL_VP	data16_13TeV	deriv	PHYS	10.0	1.86e+10	finished	620760.0
	2024-02-18 23:37:50	SWT2_CPB	mc21_13p6TeV	deriv	PHYS	3.0	1.66e+10	finished	303960.0
	2024-04-22 08:57:48	CERN	mc21_13p6TeV	deriv	PHYS	1.0	3.49e+09	failed	3300.0
	2024-03-24 17:48:13	BNL	mc20_13TeV	deriv	EGAM1	8.0	5.22e+10	finished	7010880.0
	2024-01-07 09:39:54	ANALY_ARNES_DIRECT	data18_13TeV	deriv	PHYS	1.0	2.59e+09	finished	45000.0

(b)



[1] Park, David K., et al. "AI Surrogate Model for Distributed Computing Workloads." arXiv preprint

Brookhaven National Laboratory

Generative Models for Tabular Data

Number of data – Train: 1,343,792 (60%) / validation: 447,931 (20%) / test: 447,931 (20%)

creationdate	computingsite	workload	jobstatus
2024-03-11 08:43:26	TRIUMF	244150.0	finished
2024-02-12 06:51:24	AGLT2	0.0	closed
2024-02-11 11:42:23	BNL	351720.0	finished
2024-03-17 22:52:56	TOKYO	5460.0	failed
2024-01-21 18:17:05	ANALY_ARNES_DIRECT	1173400.0	finished
2024-05-05 20:15:07	SWT2_CPB	263880.0	finished
2024-02-05 08:44:23	praguelcg2	122220.0	finished
2024-05-27 08:21:09	FZK-LCG2	185640.0	failed
2024-03-24 15:59:45	UKI-NORTHGRID-MAN-HEP	436920.0	finished
2024-04-29 03:11:47	INFN-LECCE	182300.0	finished

Samples of training data

→ Model →

creationdate	computingsite	workload	jobstatus
1.710744e+09	IN2P3-LAPP	4.775945e+04	finished
1.710744e+09	TRIUMF	1.661405e+04	finished
1.711332e+09	CERN	2.614423e+03	finished
1.714942e+09	SWT2_CPB	6.659398e+03	finished
1.713719e+09	TRIUMF	1.020332e+05	finished
...	...	...	...
1.713725e+09	NSC	8.748761e+05	finished
1.714943e+09	SWT2_CPB	3.329313e+06	finished
1.708938e+09	SWT2_CPB	1.212568e+03	finished
1.708937e+09	CERN-T0	0.000000e+00	closed
1.714940e+09	BNL	4.665673e+03	failed

synthetic data

Baselines: tabular generative models

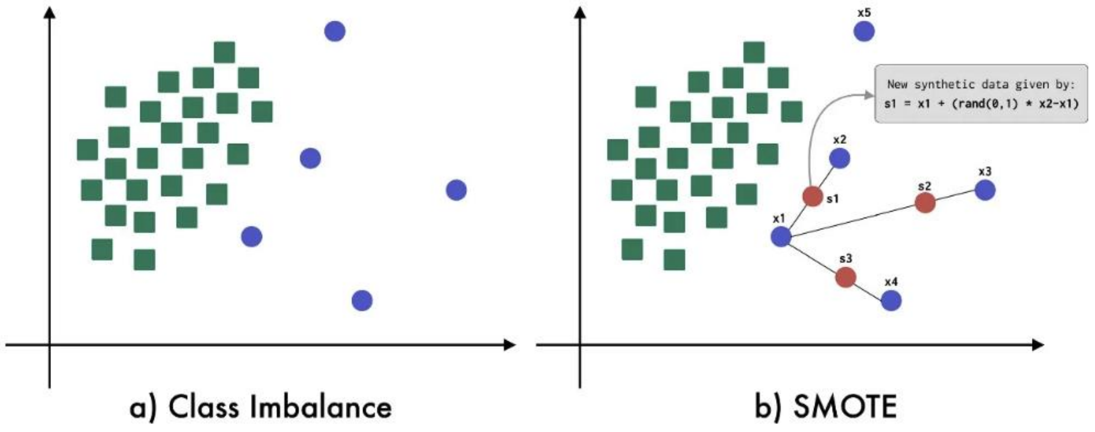
SMOTE: Non-DL algorithm working based on nearest neighbor.

TVAE: Variational autoencoder as backbone

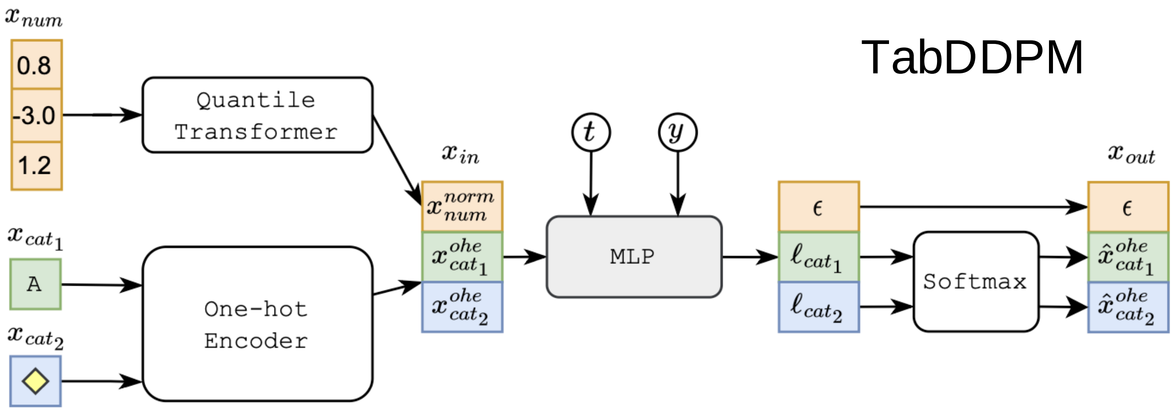
CTABGAN+: best tabular model with generative adversarial networks

TabDDPM: Diffusion model backbone

SMOTE

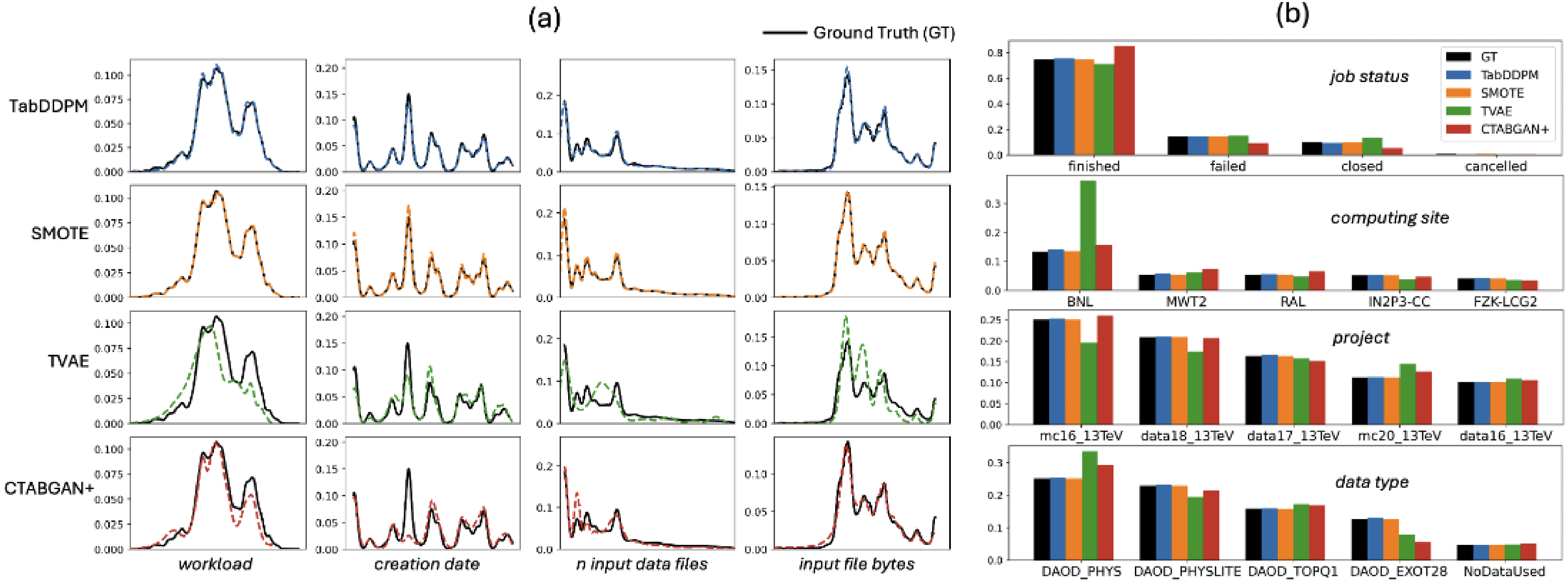


TabDDPM



Measuring Generative Performances: Results

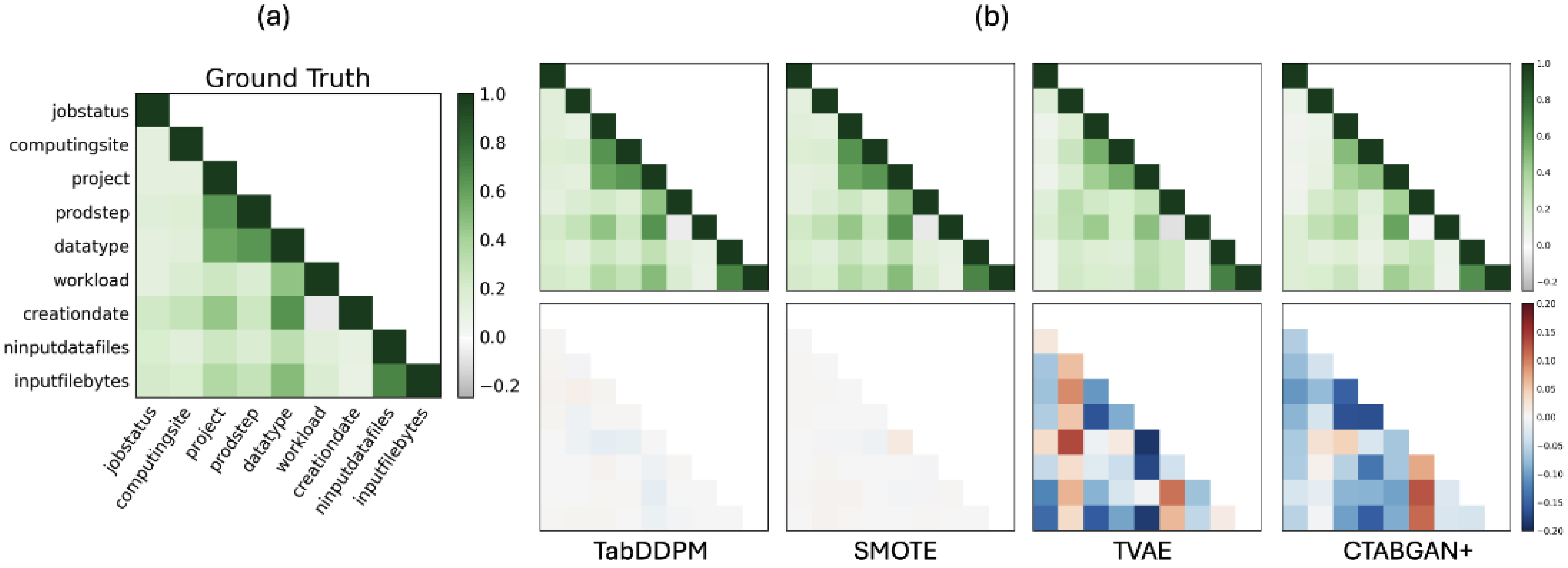
(1) Per-feature evaluation



Brookhaven National Laboratory

Measuring Generative Performances: Results

(2) Correlations between feature pairs



Measuring Generative Performances: Results

(3) Minimizing privacy risk: distance to closest record (DCR)

TABLE I
PERFORMANCE COMPARISONS ON SURROGATE MODELS.

Model	WD ↓	JSD ↓	diff-CORR ↓	DCR ↑	diff-MLEF ↓
TVAE	0.961	0.806	0.653	**0.143**	5.875
CTABGAN+	1.0	0.820	0.658	0.105	10.464
SMOTE	**0.871**	**0.799**	**0.011**	0.001	**0.058**
TabDDPM	0.874	**0.799**	0.036	0.025	0.826

Implementation overview

Input

To optimize the dispatcher in a real scenario, we need real data, or a surrogate model of the data. We take 5-months WMS PanDA job records as our real data for which we build AI surrogate models.

Outcome

Summarized by several metrics, such as queue time, error rate, or data movement, providing an introspective performance measures of the policy.

Dispatcher

Our objective is to build a centralized dispatcher which steers both job scheduling and dataset movement, contributing to the resilience of the workflow system.
Working on reinforcement learning based model as a dispatcher for jobs and datasets.

Environment

To optimize the dispatcher, we also need a **realistic environment of distributed computing** facilities, which takes jobs and datasets as inputs and produce estimated quality performances. **Considering SimGrid, WRENCH, DCSim as potential programs in simulation.**

Brookhaven National Laboratory

Conclusion

- Curated and analyzed 150-day WMS PanDA records.
- Identified key performance metrics and representative columns.
- Built AI surrogate model for the PanDA records [1]. The surrogate model successfully learns the joint distribution of WMS PanDA table as well as the time dynamics.
- Future work includes incorporating more diverse features of PanDA, developing simulated distributed computing environment, and a dispatcher optimized for resiliency.

[1] Park, David K., et al. "AI Surrogate Model for Distributed Computing Workloads." arXiv preprint