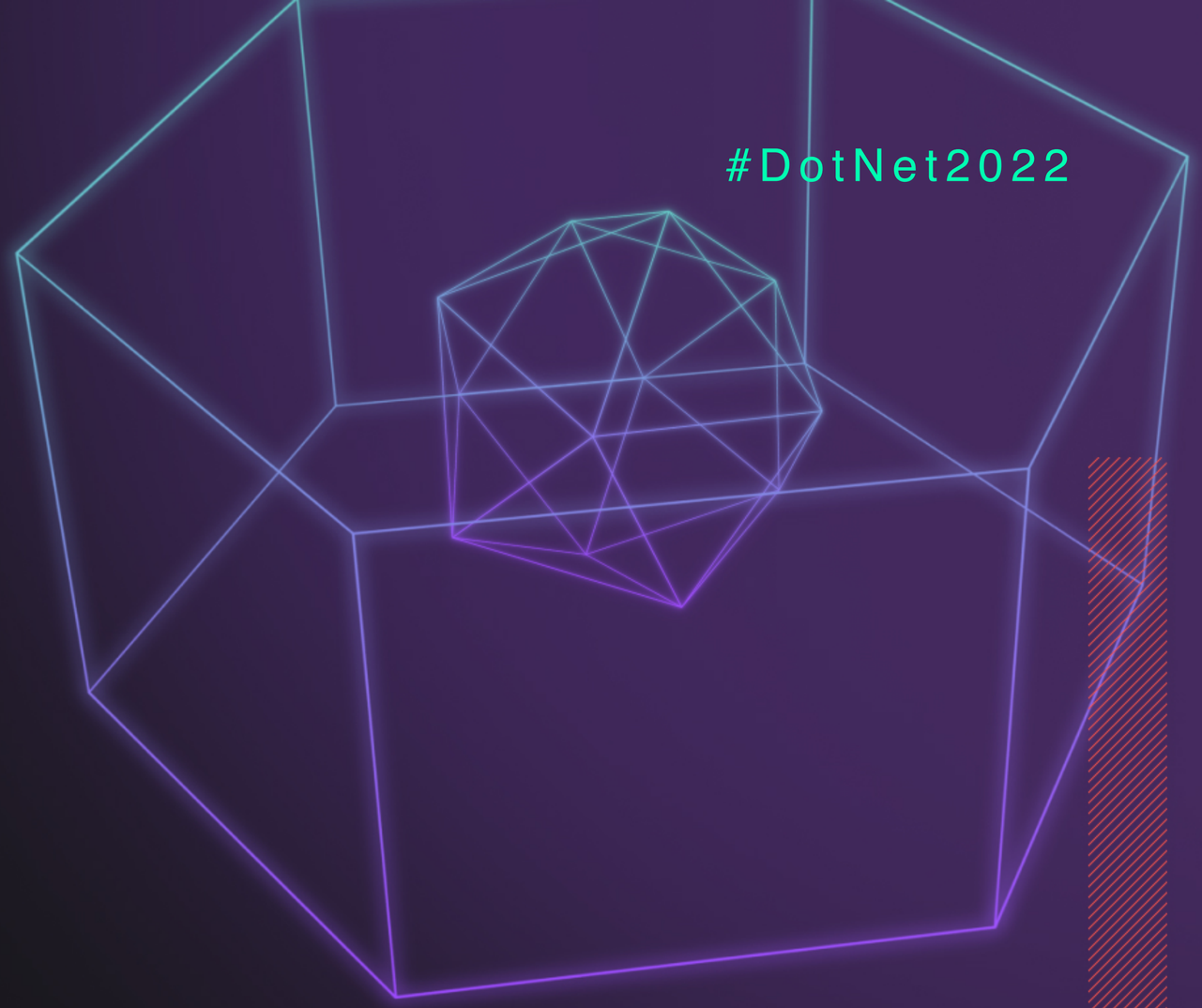
DotNet 2022

TECH CONFERENCE

#DotNet2022



28th June

Let’s focus more on data !

Powered by

plain concepts

DotNet 2022

SPONSORS

Microsoft

intel

intelequia

LEMON CODE

DevsDNA

COLLABORATORS

<XantarDev/>
/* Technical Community */

My Public Inbox

SEVILLADOTNET
COMUNIDAD .NET

Power Platform Valencia

W4TT

#DotNet2022

DotNet2022

TECH CONFERENCE

#DotNet2022

Alexander González

Data Scientist / ML Engineer

Tech lover🚀

Data and AI I Computer Vision🧠 📸 🌄

Microsoft Artificial Intelligence MVP🏆

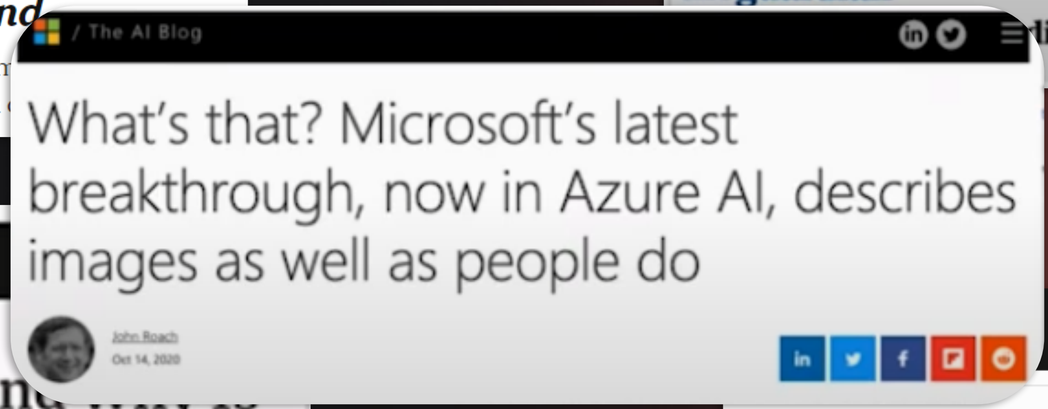
@alexndrglez

alexglezglez96@gmail.com

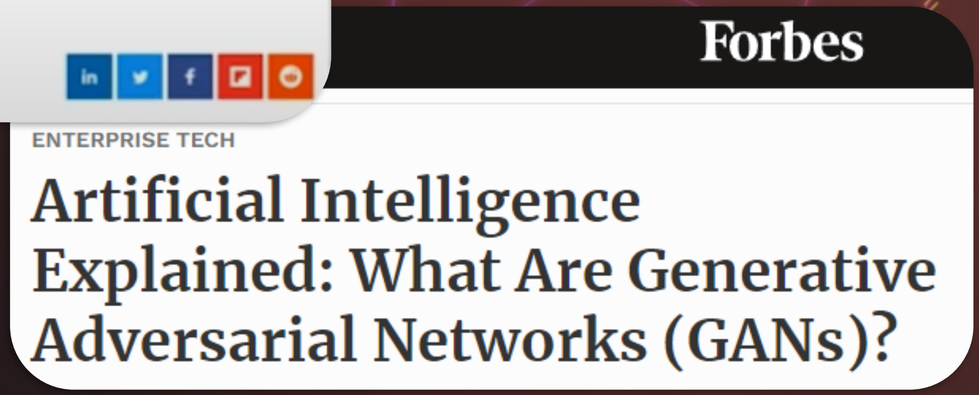
AI Success


The New York Times
Meet DALL-E, the A.I. That Draws Anything at Your Command
New technology that blends language and im
graphic artists — and speed disinformation

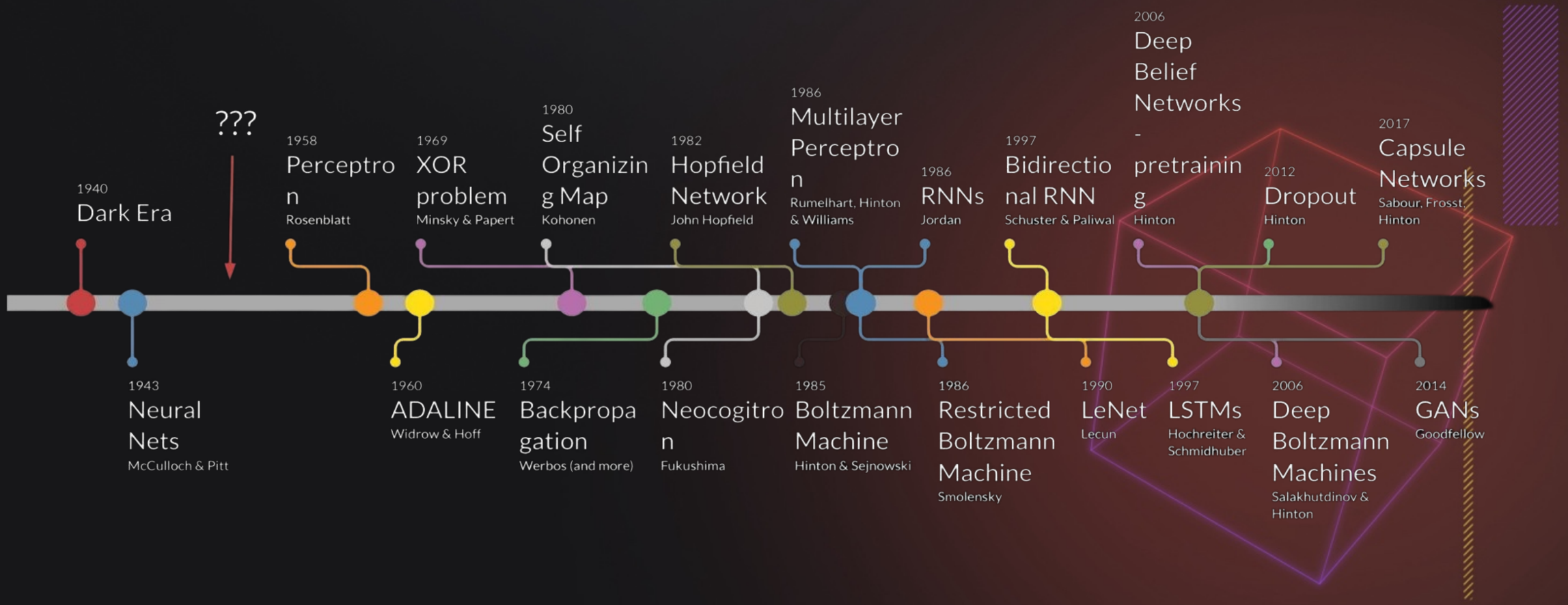

Google's DeepMind predicts 3D shapes of proteins
theguardian
ding of proteins could usher in new era of


/ The AI Blog
What's that? Microsoft's latest breakthrough, now in Azure AI, describes images as well as people do
John Roach
Oct 14, 2020

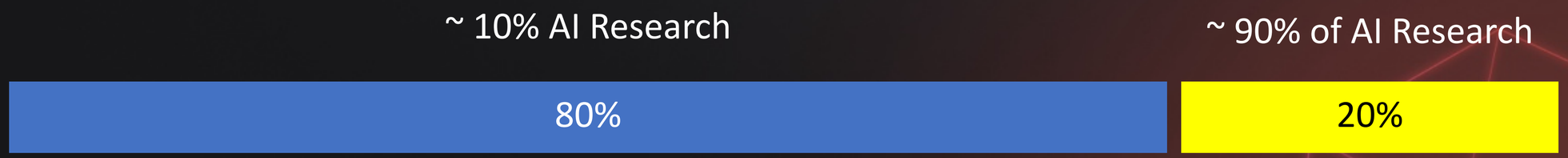

Forbes
21,547 views | Oct 5, 2020, 12:21am EDT
What Is GPT-3 And
It Revolutionizing Artificial Intelligence?
Bernard Marr Contributor
Enterprise Tech


Forbes
ENTERPRISE TECH
Artificial Intelligence Explained: What Are Generative Adversarial Networks (GANs)?

Deep Learning Timeline



AI Frameworks History



Datasets Augmentation?
Datasets Comparisons?
Datasets Collection and Manipulation?
Datasets Optimal Transport?
Datasets Labeling?
Datasets techniques to handle noisy data?

AI system = Code + Data

(model / algorithm)

AI Status-Quo

Model-Centric

Fixed data, adapt model

Learning focuses on adapting model parameters to suit new datasets. Collect what data you can and develop a model good enough to deal with the noise in the data

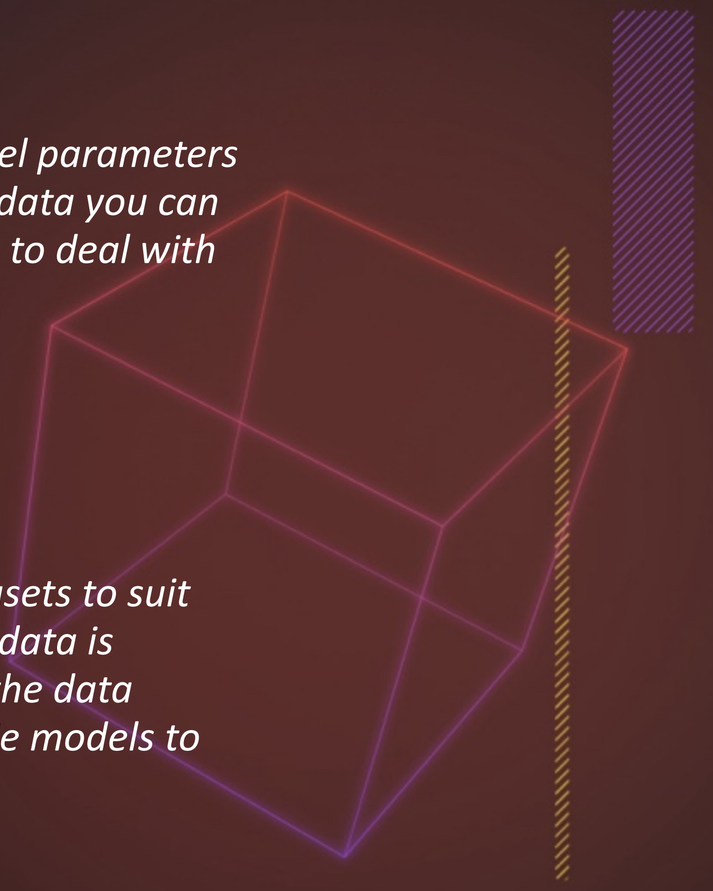
Iteratively improve Code/Model

Data-Centric

Fixed model, adapt data

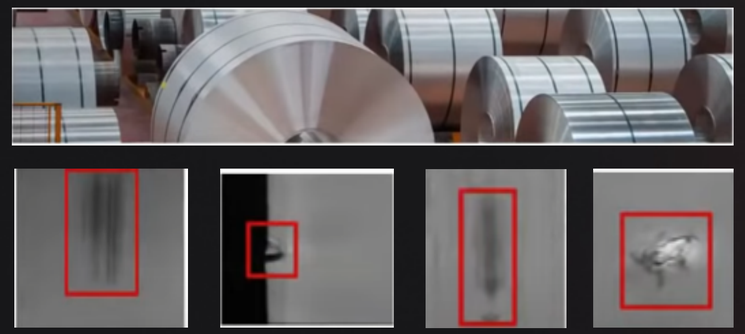
Learning focuses on adapting datasets to suit the model. The Consistency of the data is paramount. Use tools to improve the data quality. This will way allow multiple models to do well.

Iteratively improve the data



Data-Centric AI – Label Consistency

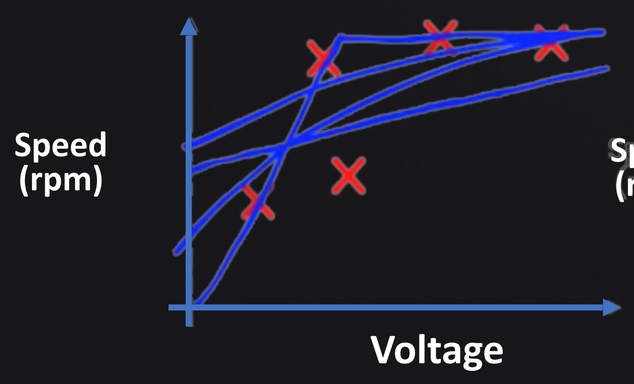
Examples defects



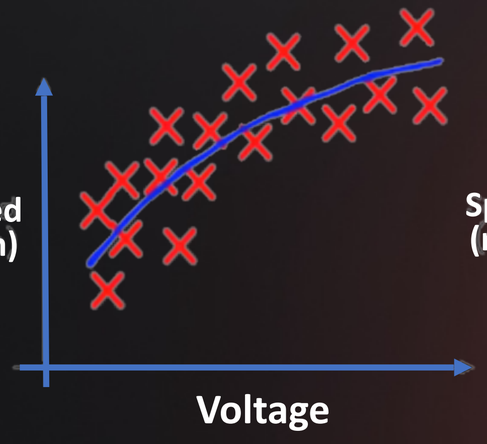
Baseline system: 76.2% accuracy
Target: 90% accuracy

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

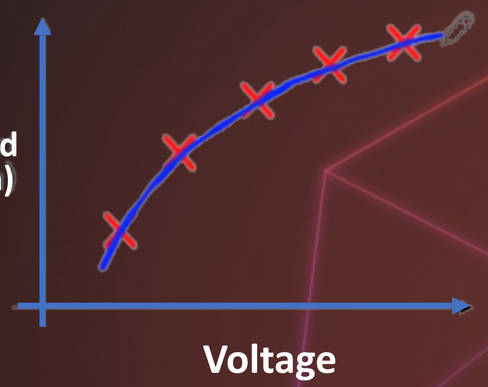
Data-Centric AI – Label Consistency



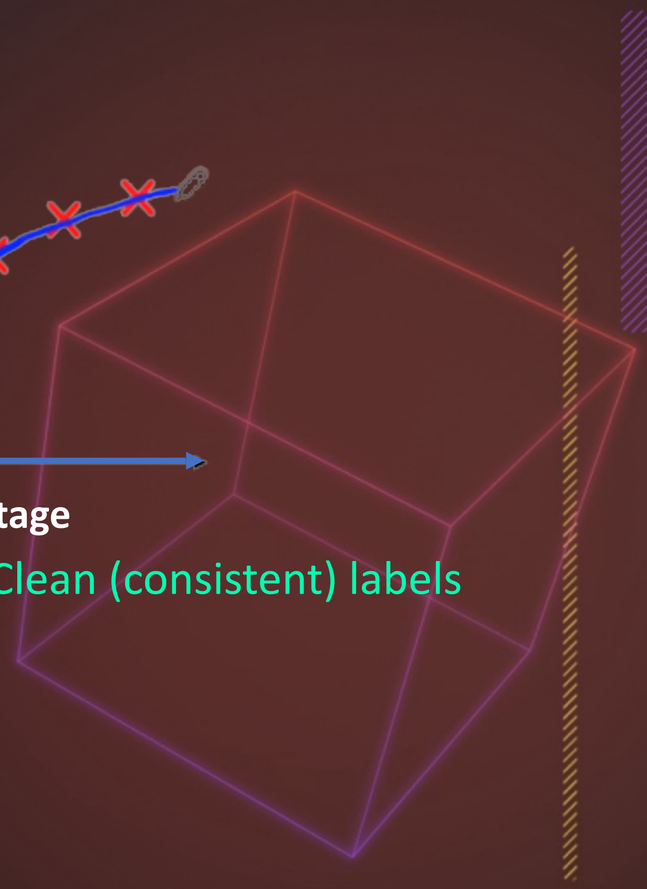
Small data, Noisy labels



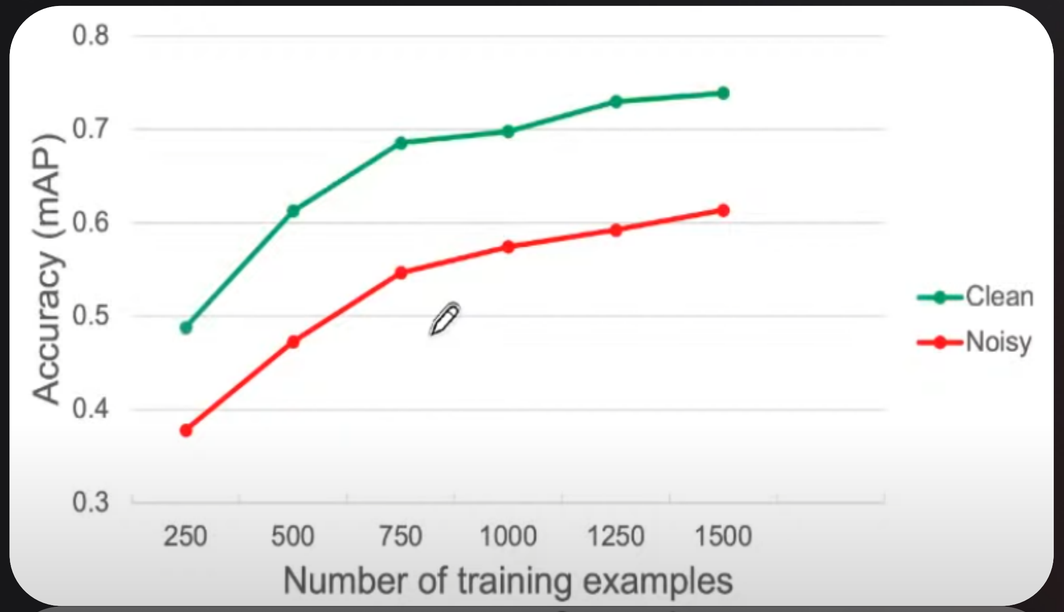
Big data, Noisy labels

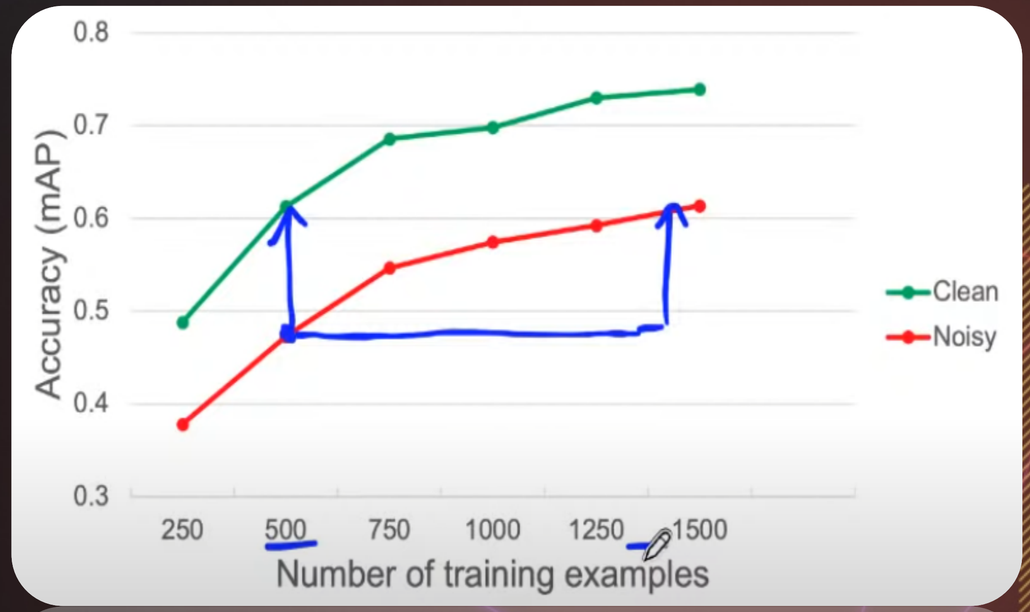


Small data, Clean (consistent) labels



Data-Centric AI – Label Consistency





Data-Centric AI – Datasets scarcity

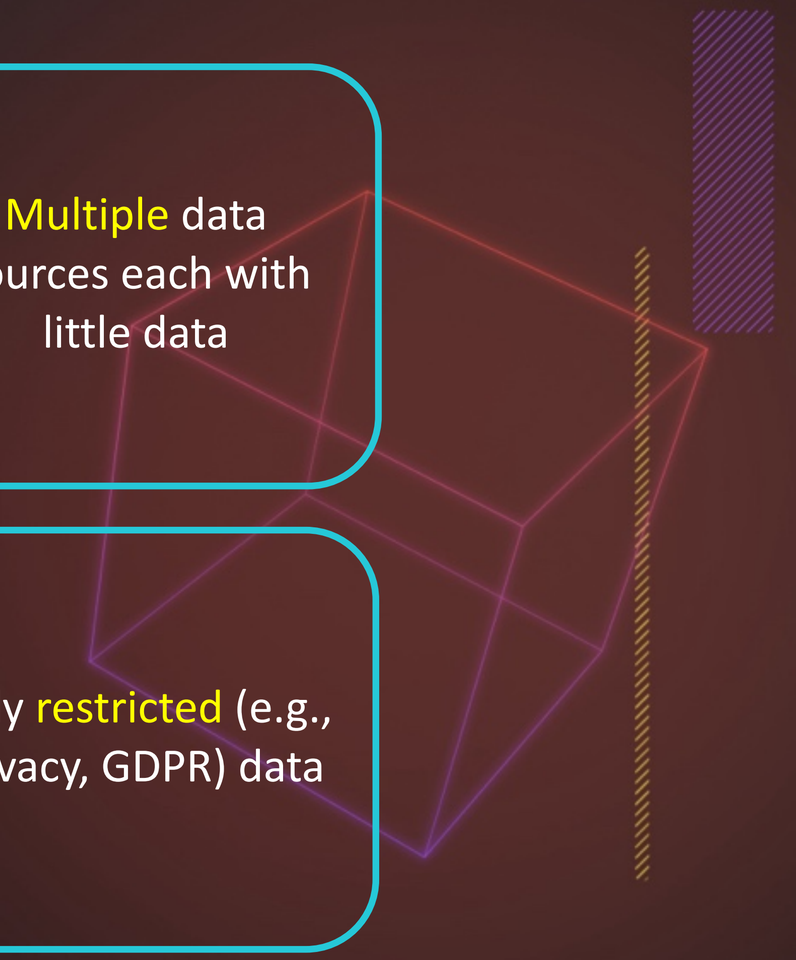
No data, but data from similar domain available

Labeled data scarce, unlabeled data available

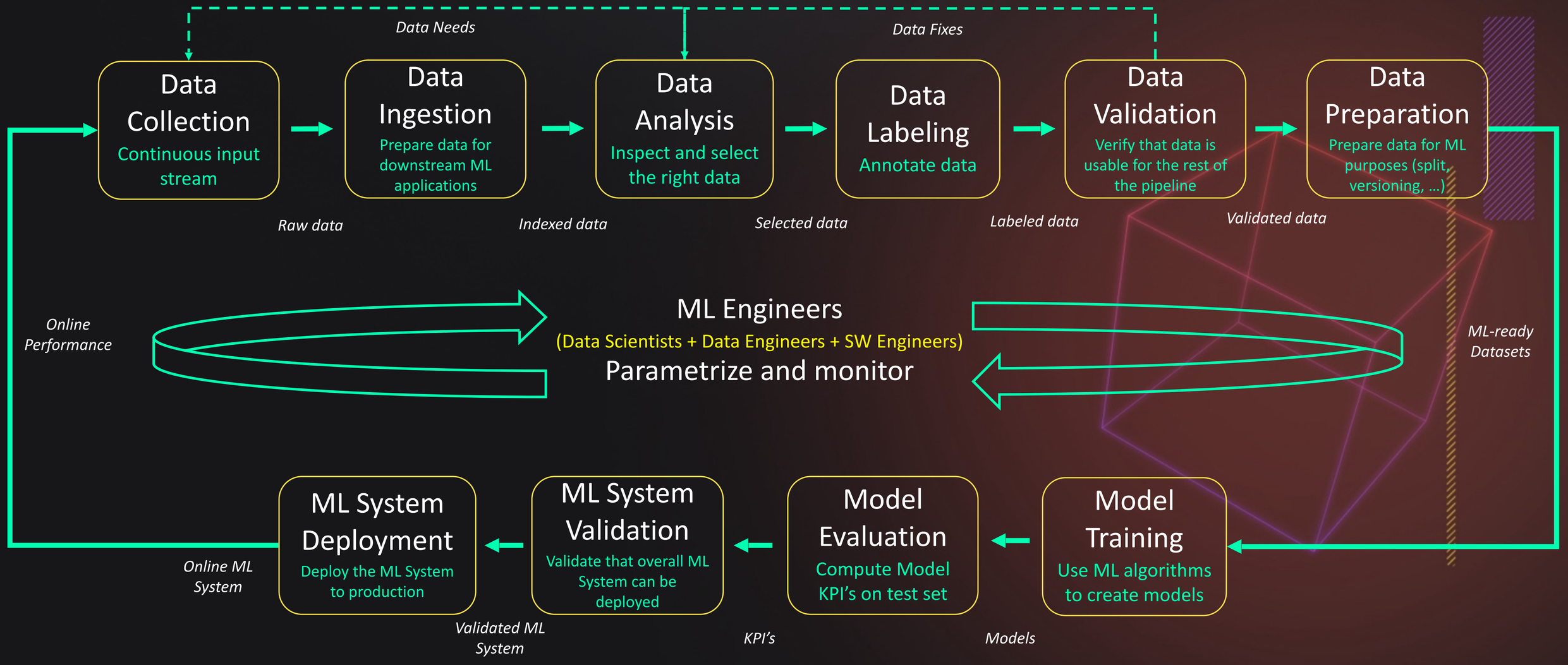
TARGET DATASET SCARCITY

Multiple data sources each with little data

Only restricted (e.g., privacy, GDPR) data



Data-Centric AI - MLOps



Data-Centric AI – Model as a service

Current AI: model adapted to data

Model Provider

Client / Tenant

Data-Centric AI: data adapted to model

Current AI

Requires fine-tuning model (time / cost & model size)

Model shipped to client (proprietary info loss: model, data)

Data Shipped to provider (privacy, data protection issues)

Client data enriches model (potential legal issues)

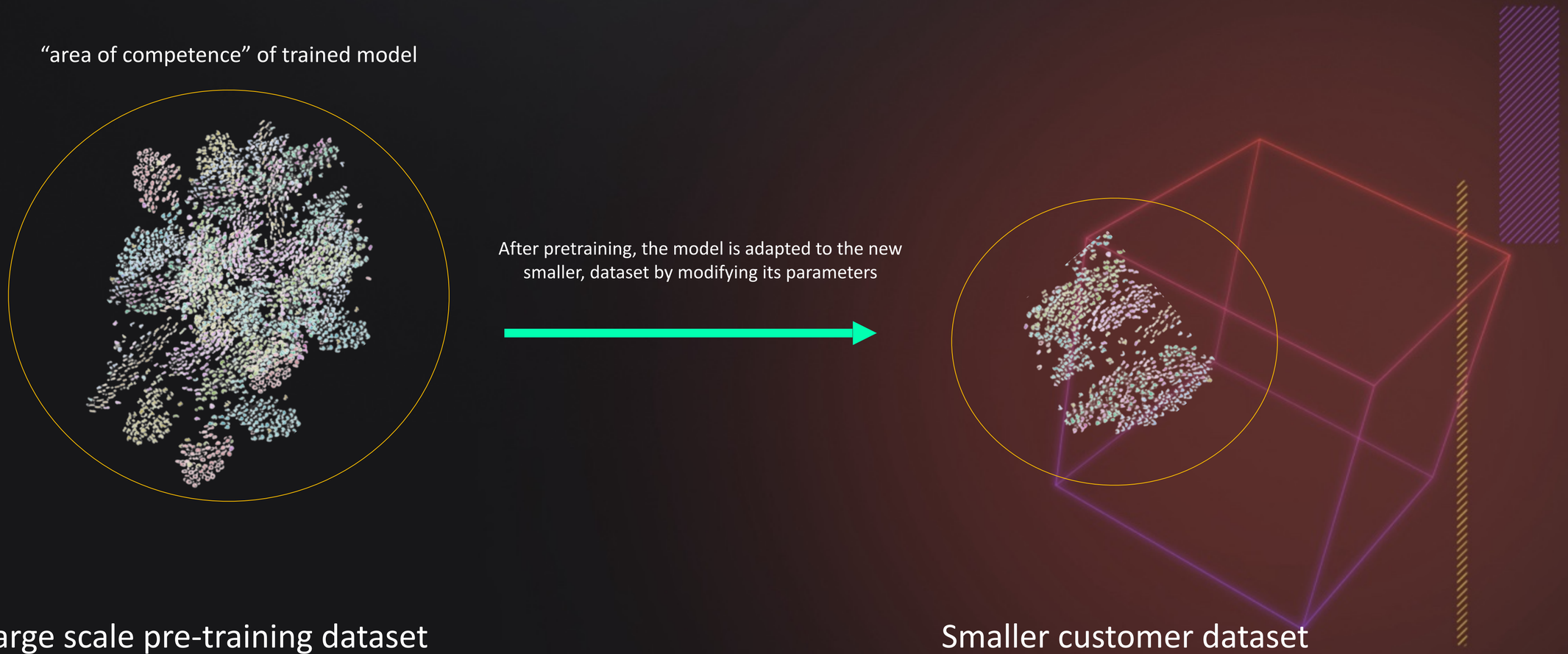
Data-Centric AI

Model is not changed

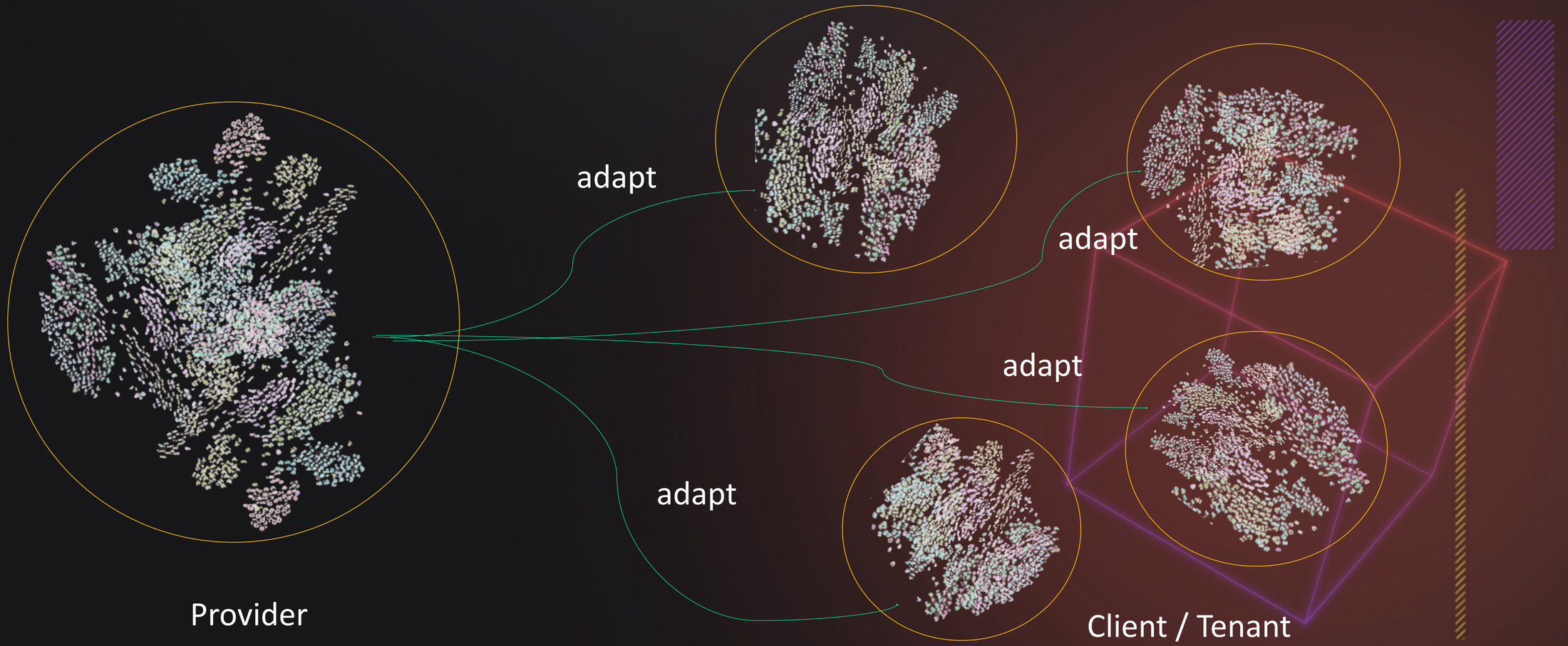
Model stays with provider, data stays with client

Model probably not enriched with the client data

Data-Centric AI – Model Pre-training/Transfer Learning



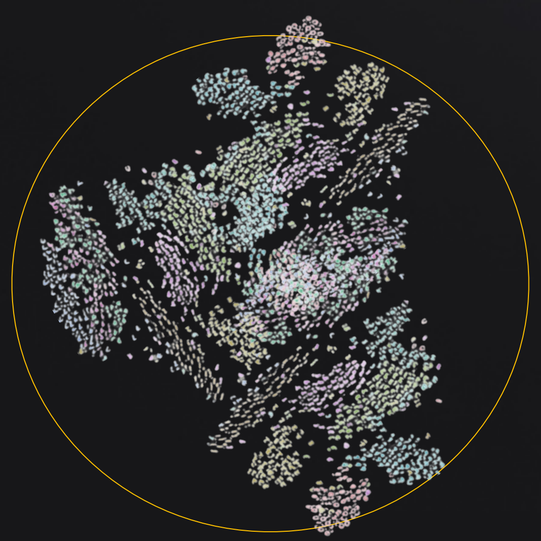
Issue 1: Computation and storage cost

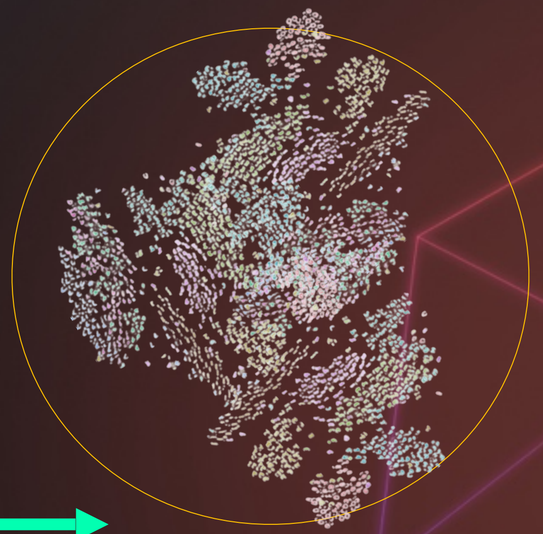


Issue 2: Architecture / data bleed

Proprietary architecture

Private training data





Access to architecture specs

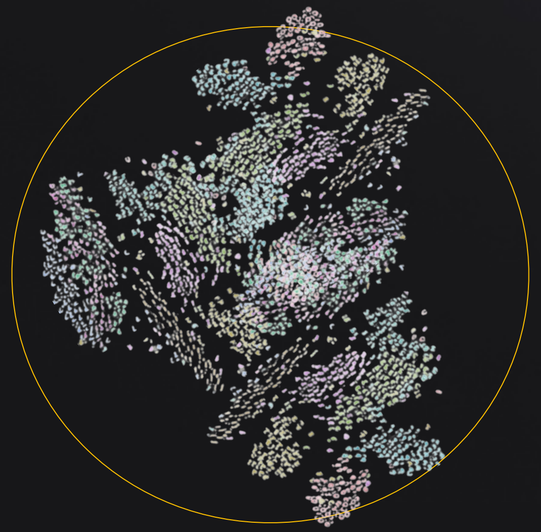
Adversarial queries (recover train data instances)

Provider side

Client side

Issue 3: Client-side model retention

Proprietary architecture



Provider side

Privacy / Compliance Boundary

Model enriched with client data

Private customer data

Client side

Solution - Learning + Generating datasets

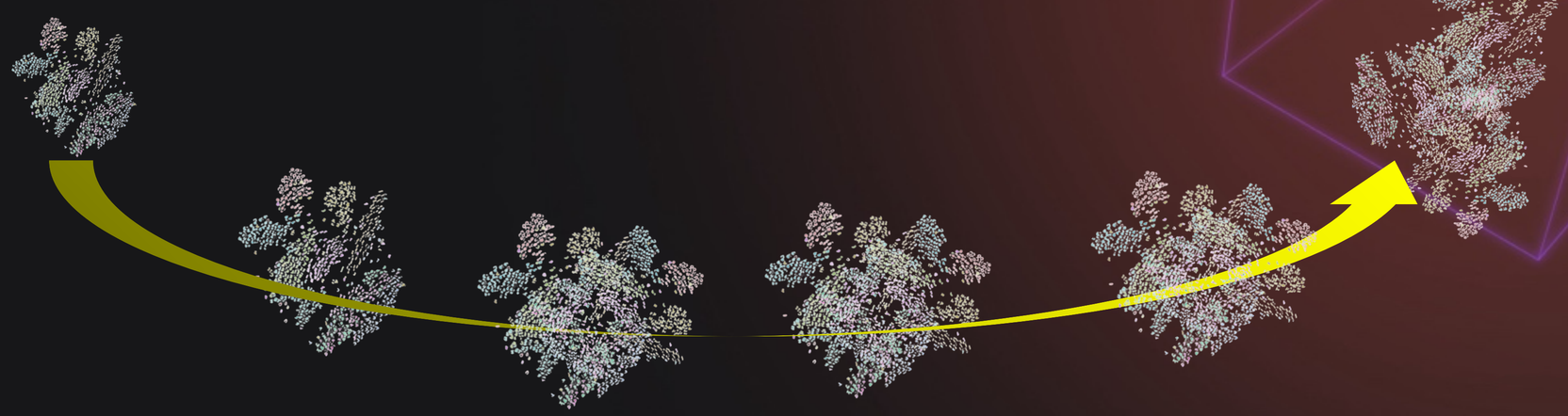
Similarity to provider's pretrain dataset

Gradient Flow using optimal transport

Initial dataset + objective + constraints → Learned dataset

e.g single customer data

Private attributes not identifiable (e.g differential privacy)

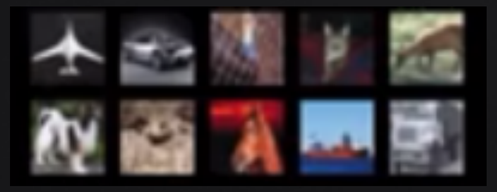


Data-Centric AI

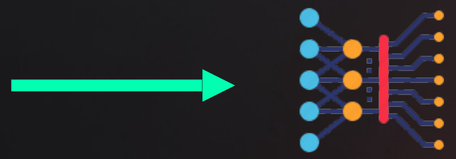
Addresses issues of Model-Centric AI

- ✓ *Reduces computations / storage cost for large-scale model serving*
- ✓ *Mitigates architecture / data bleed*
- ✓ *Prevents client-side model retention*

Teaser – Models re-purposing

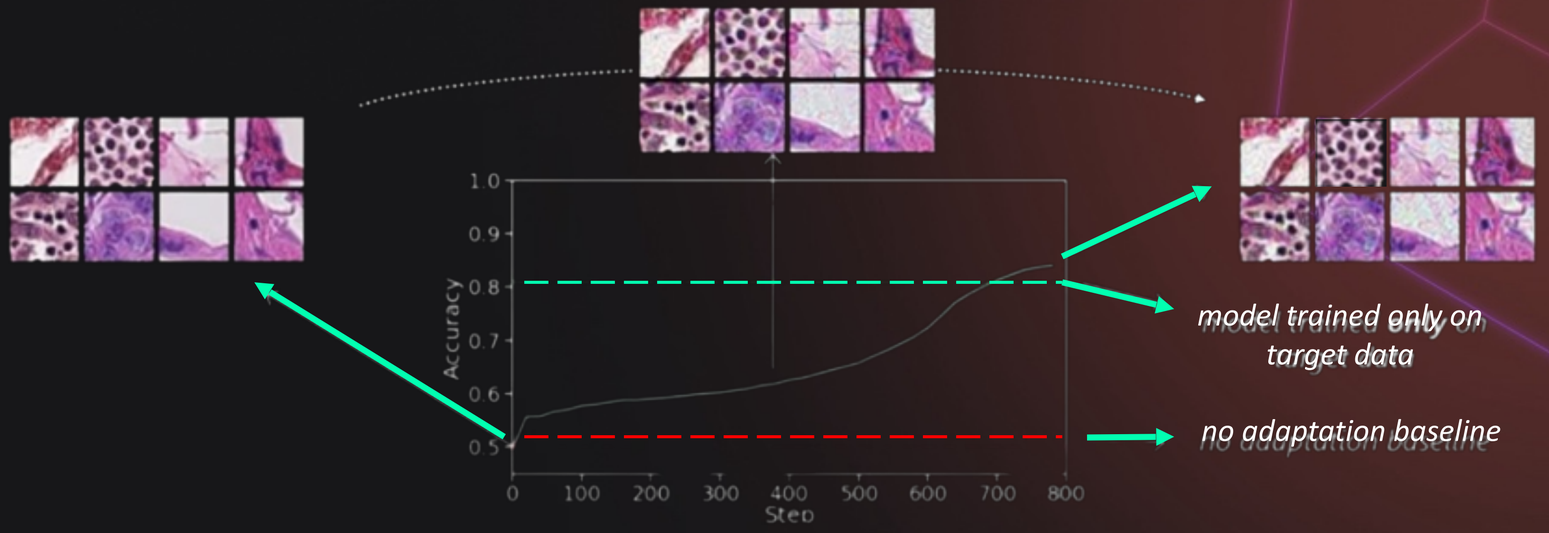


Model: *e.g CIFAR-10 Dep Neural network trained to recognize common objects (frozen)*



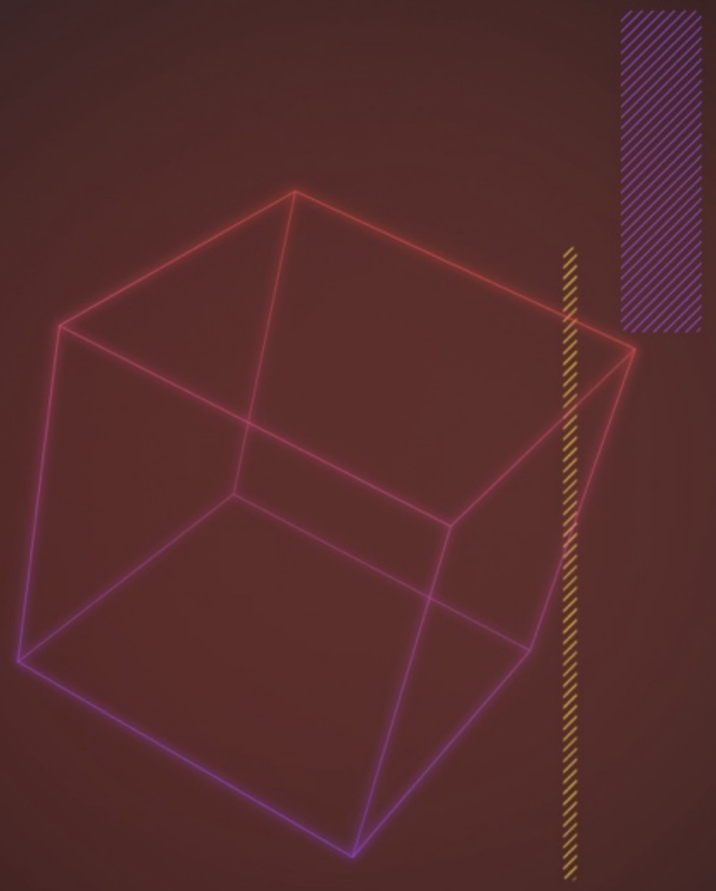
Classification model

Target Dataset: *H&E images of lymph nodes section from breast cancer patients*



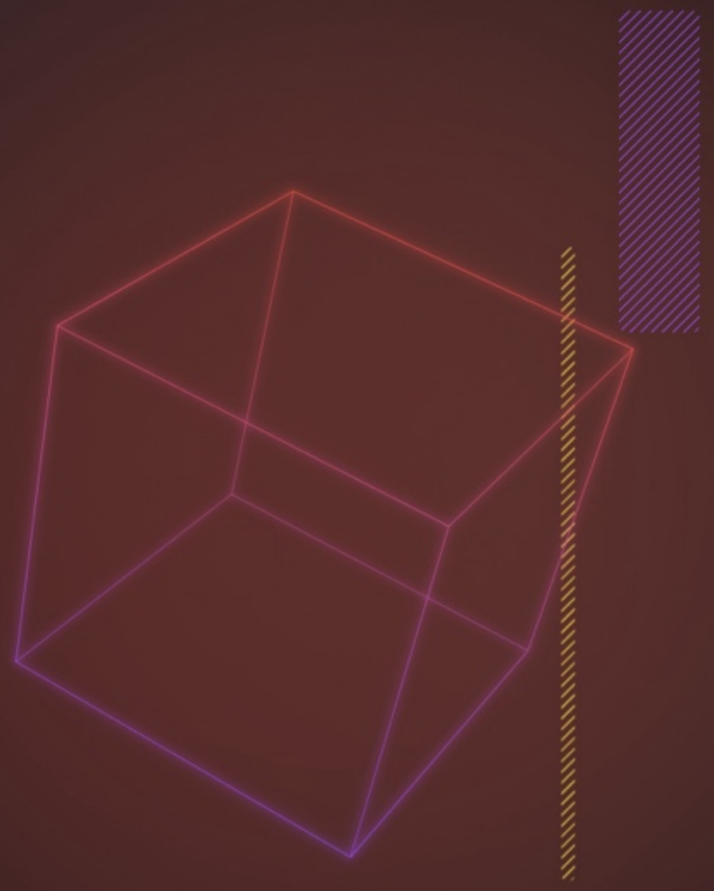
Technical Content. Don’t be scared

- **Optimal Transport (OT)**
  [Gaspar Monge (1746-1818) & Kantorovich Formulation]
- **OT distances between datasets**
  [Alvarez-Melis & Fusi, NeurIPS 2020]
- **Gradient Flows between datasets**
  [Alvarez-Melis & Fusi, ICML 2021]



Optimal Transport

A flexible geometric method for comparing probability distributions, and can be used to compare *any two datasets*, regardless of whether their label sets are directly comparable.

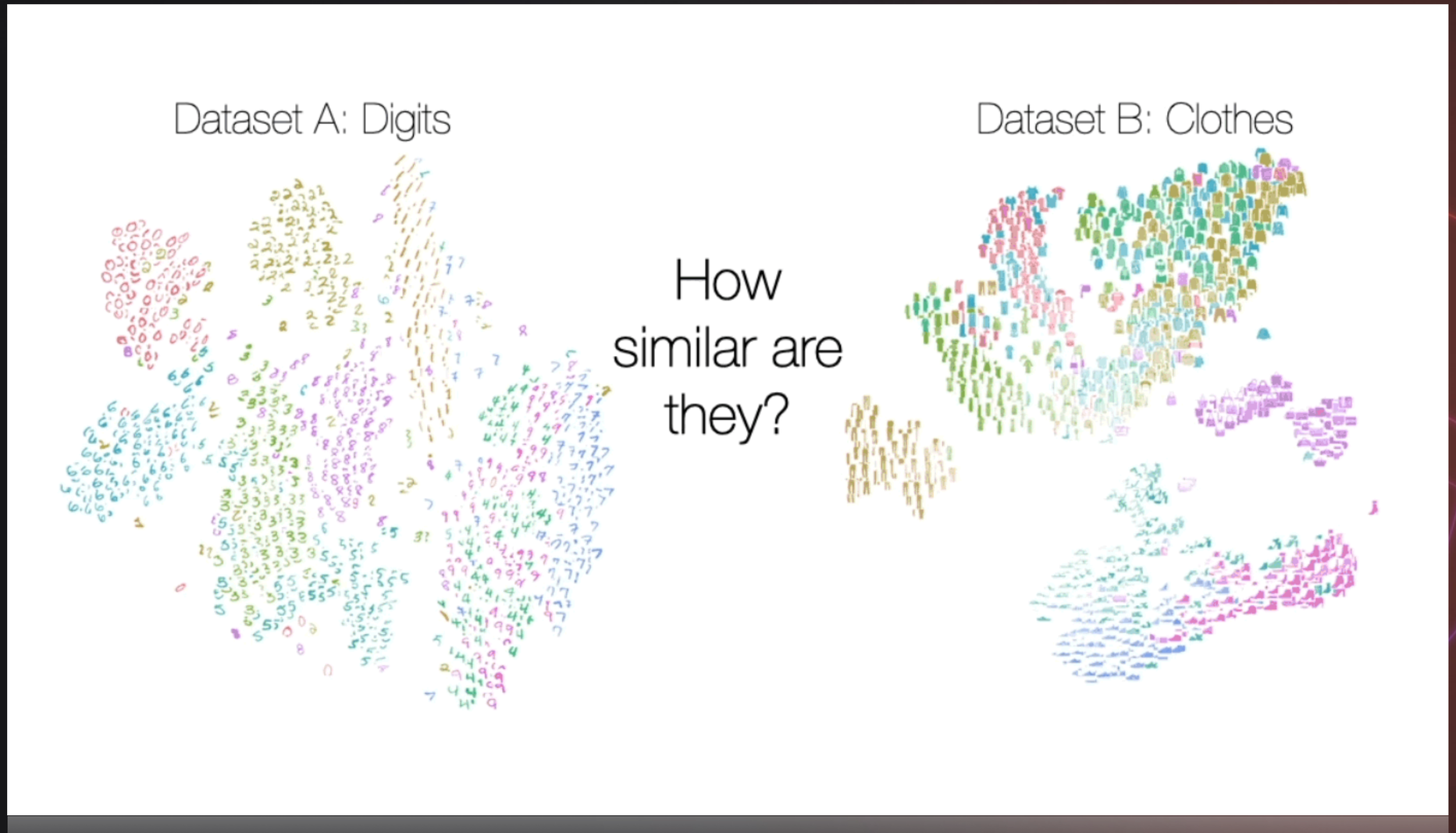


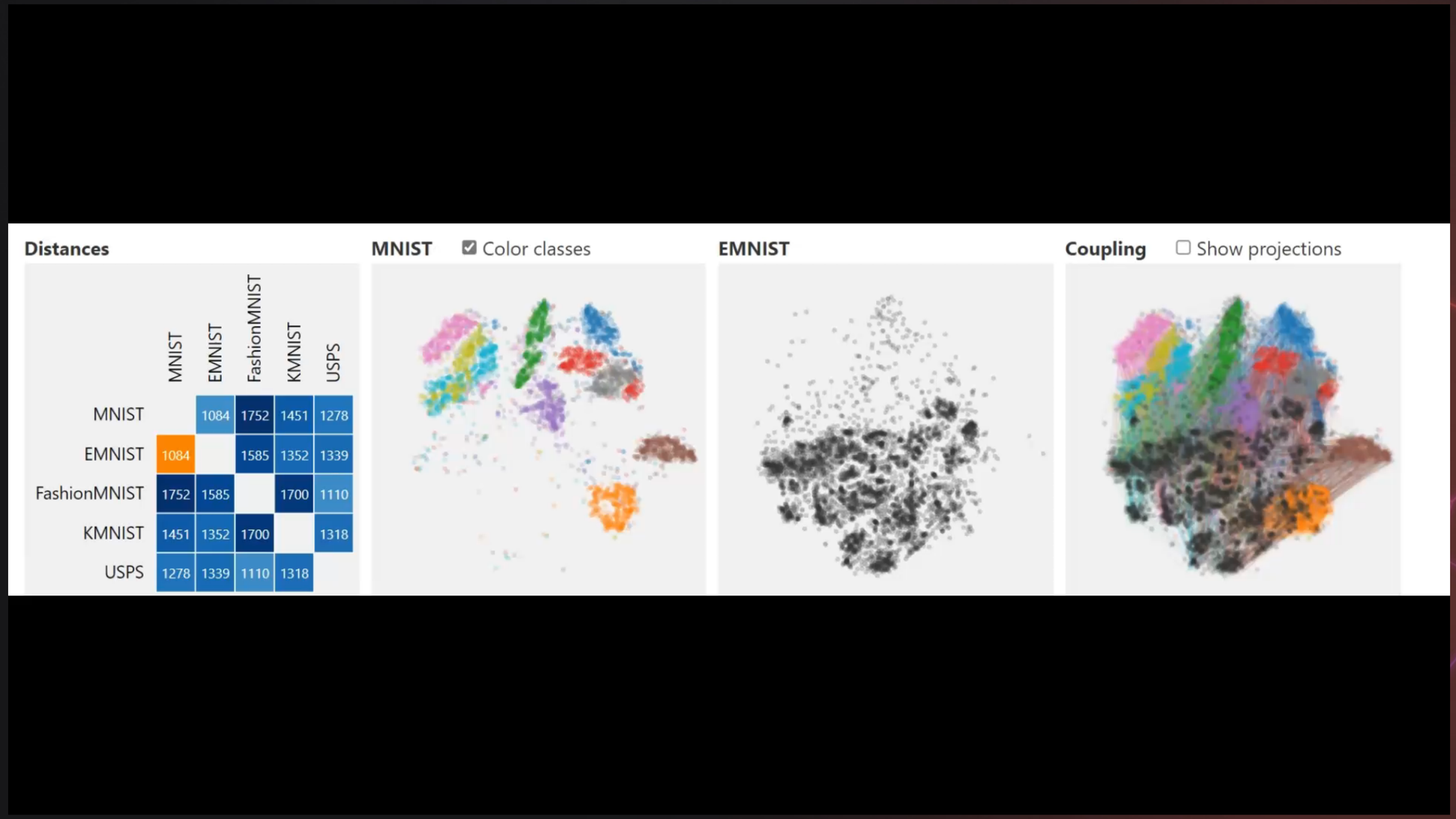
Optimal Transport Dataset Distance

An approach to defining and computing similarities, or distances, between classification datasets. The OTDD relies on optimal transport (OT).

OTDD returns a coupling of the two datasets being compared, which can be understood as a set of soft correspondences between individual items in the datasets.

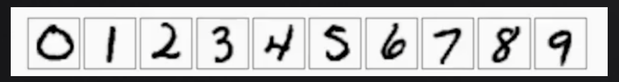
Optimal Transport - Distances between datasets




Distances
MNIST
EMNIST
FashionMNIST
KMNIST
USPS
MNIST 1084 1752 1451 1278
EMNIST 1084 1585 1352 1339
FashionMNIST 1752 1585 1700 1110
KMNIST 1451 1352 1700 1318
USPS 1278 1339 1110 1318
MNIST
Color classes
EMNIST
Coupling
Show projections

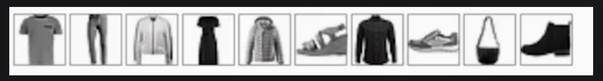
Distances between datasets – Predicting transferability

MNIST

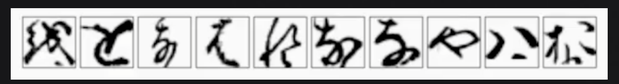


EMNIST

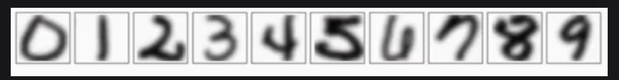
Fashion MNIST

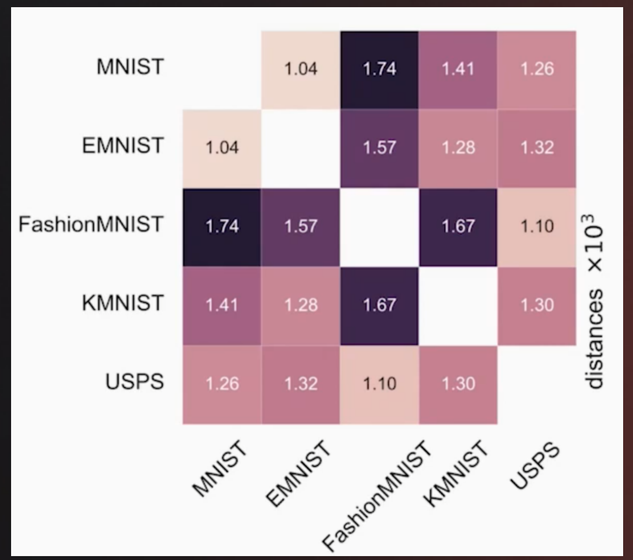


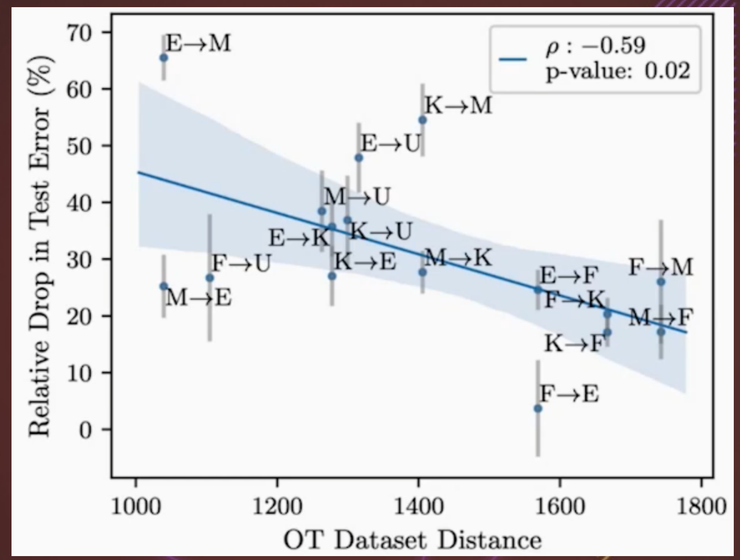
KMNIST



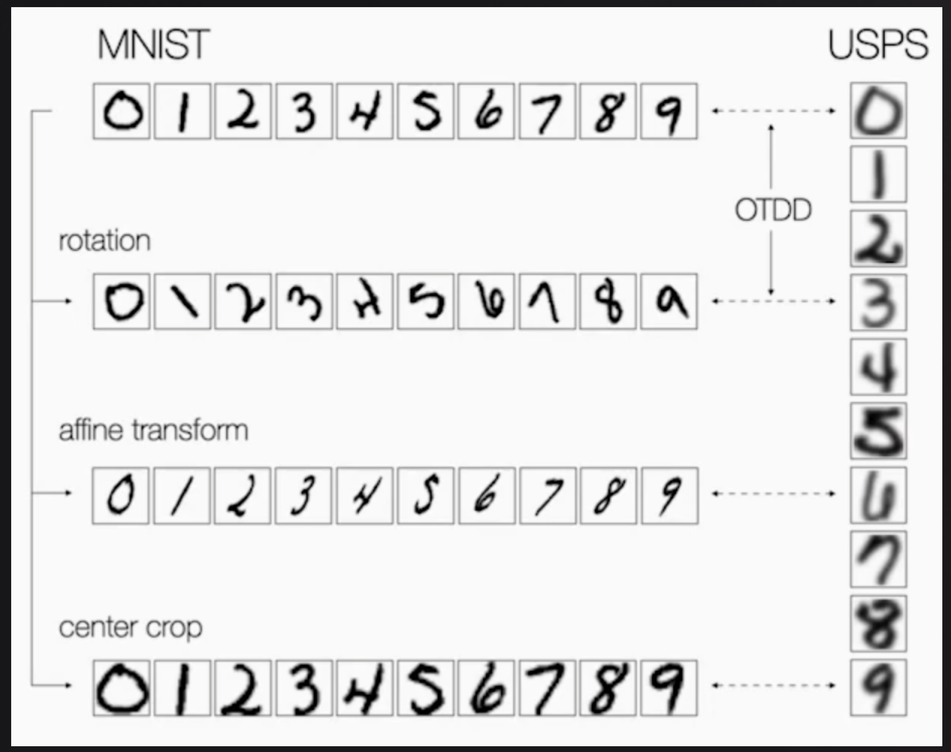
USPS

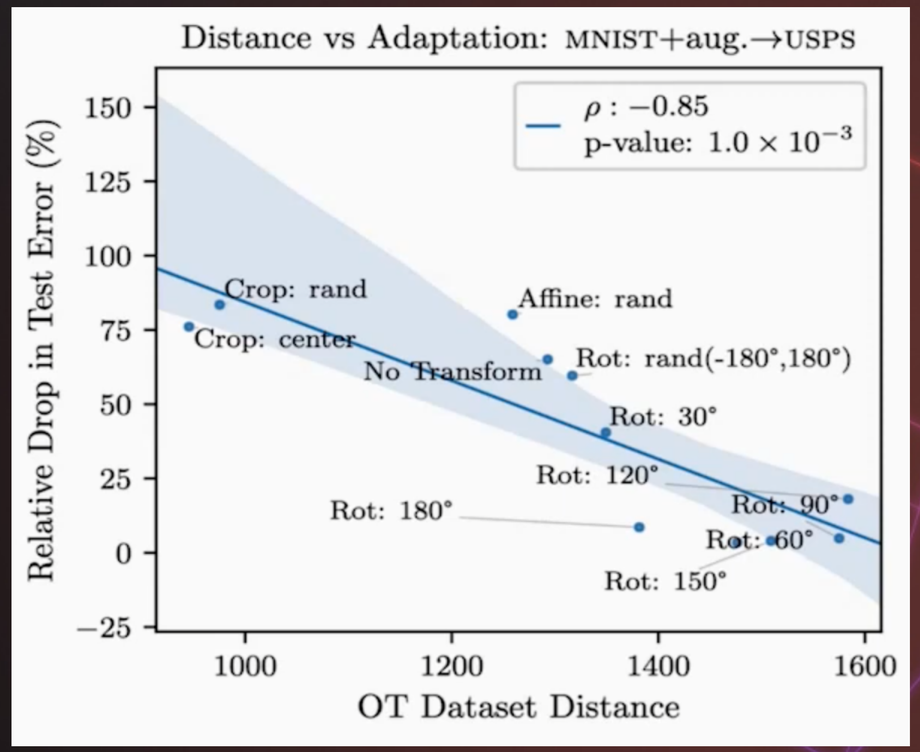




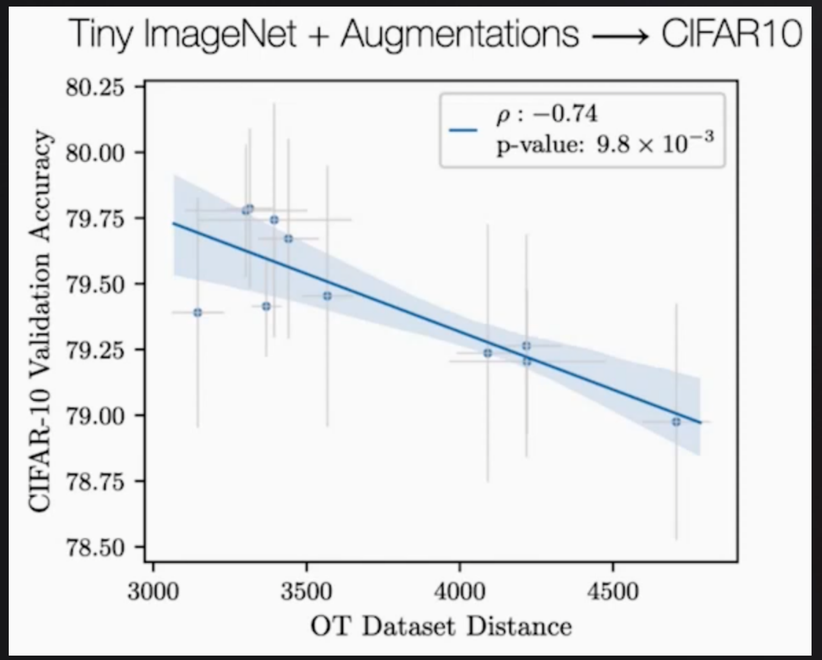


Predicting transferability – Data Augmentations



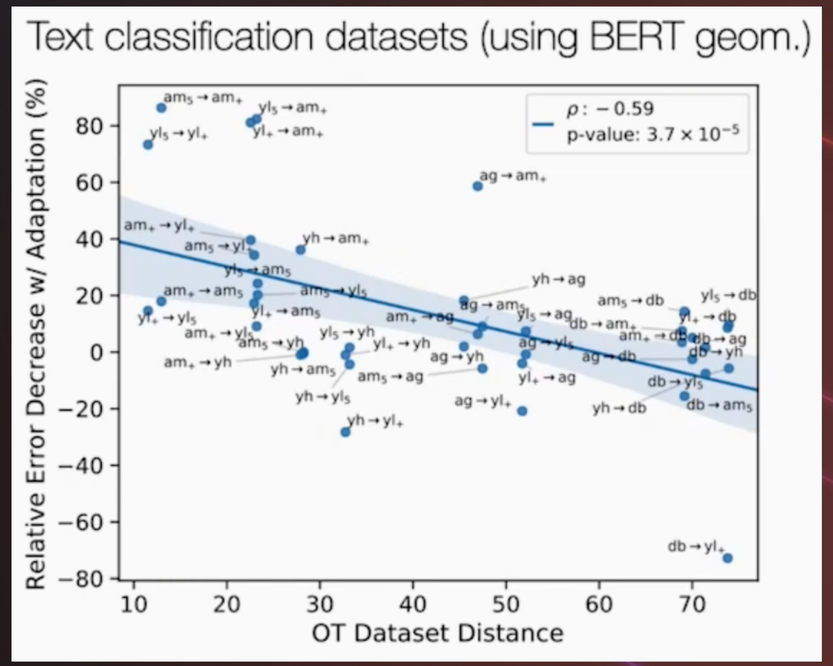


Predicting transferability – Further results



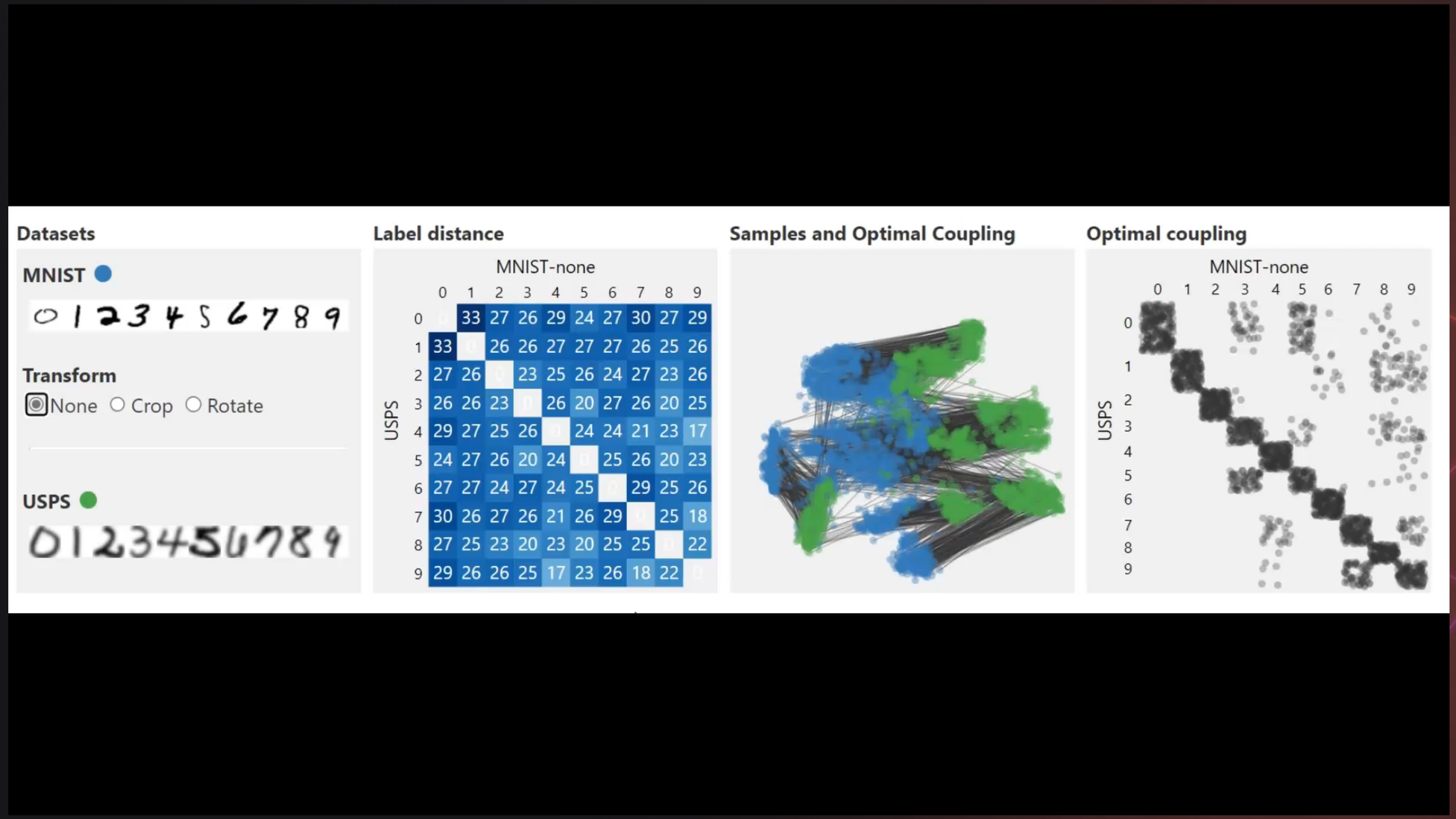
Vision classification datasets

Imagenet, CIFAR10 …



Sentence classification datasets

AG News, DBPedia, Yelp Reviews, Amazon Reviews, Yahoo Answers …

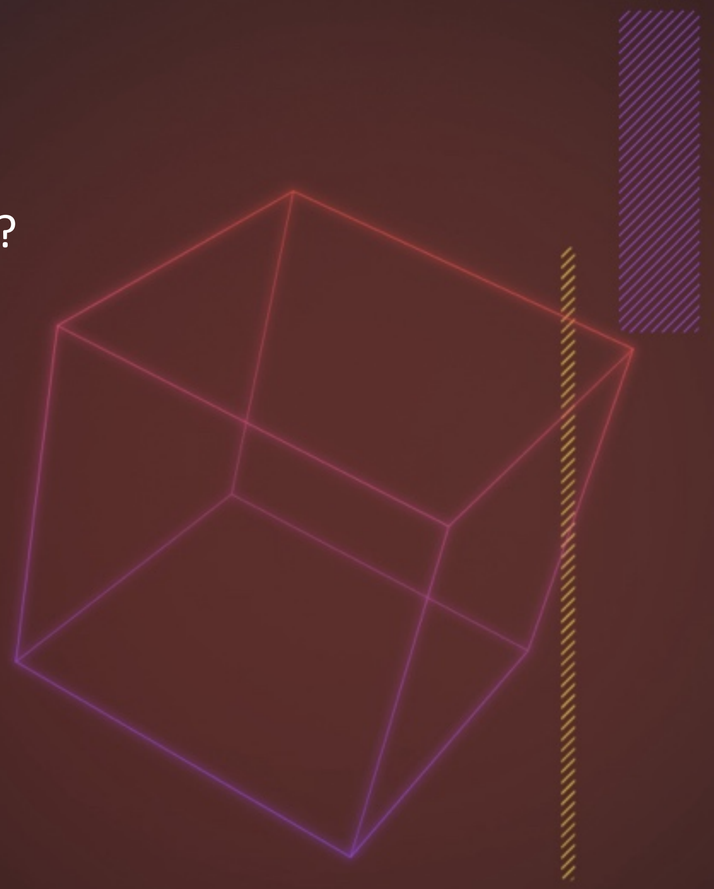

Datasets
MNIST
Transform
None
Crop
Rotate
USPS
Label distance
MNIST-none
USPS
0 1 2 3 4 5 6 7 8 9
0 33 27 26 29 24 27 30 27 29
1 33 26 26 27 27 27 26 25 26
2 27 26 23 25 26 24 27 23 26
3 26 26 23 26 20 27 26 20 25
4 29 27 25 26 24 24 21 23 17
5 24 27 26 20 24 25 26 20 23
6 27 27 24 27 24 25 29 25 26
7 30 26 27 26 21 26 29 25 18
8 27 25 23 20 23 20 25 25 22
9 29 26 26 25 17 23 26 18 22
Samples and Optimal Coupling
Optimal coupling
MNIST-none
USPS

Shaping Datasets

- So far, datasets are fixed, use OTDD to guide model adaption
- What if model is fixed can we modify datasets to minimize the distance (OTDD)?

Why Shape Datasets?

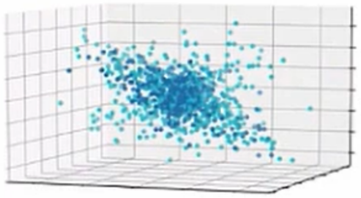
- Protect sensitive attributes
- Increase class separation
- Re-purpose an already-trained model



Shaping Datasets

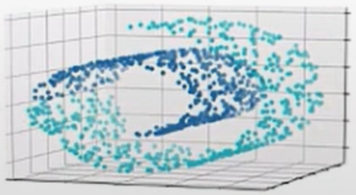
Application:
Dataset Shaping

Initial Dataset ρ_0



Want: ρ^* but linearly separable

Reference Dataset ρ^*



Functional Objective $F(\rho)$

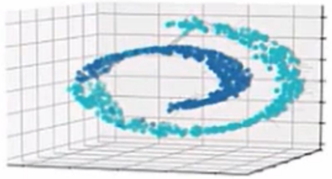
$\text{OTDD}(\rho, \rho^*)$
(distance only)

$\mathcal{V}(\rho)$
(constraint only)

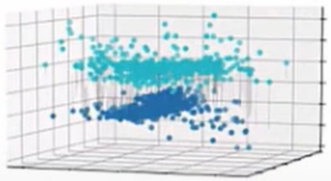
$\text{OTDD}(\rho, \rho^*) + \mathcal{V}(\rho)$
(distance + constraint)

Optimization Iterates ρ_t

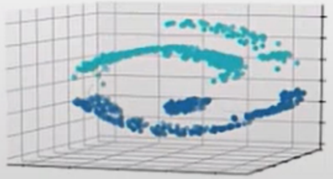
Time t=50.00, $F(\rho_t)$=2.22



Time t=15.00, $F(\rho_t)$=1.07



Time t=50.00, $F(\rho_t)$=32.72



Shaping Datasets – Gradient Flow

Initial particles

Particles after t flow steps

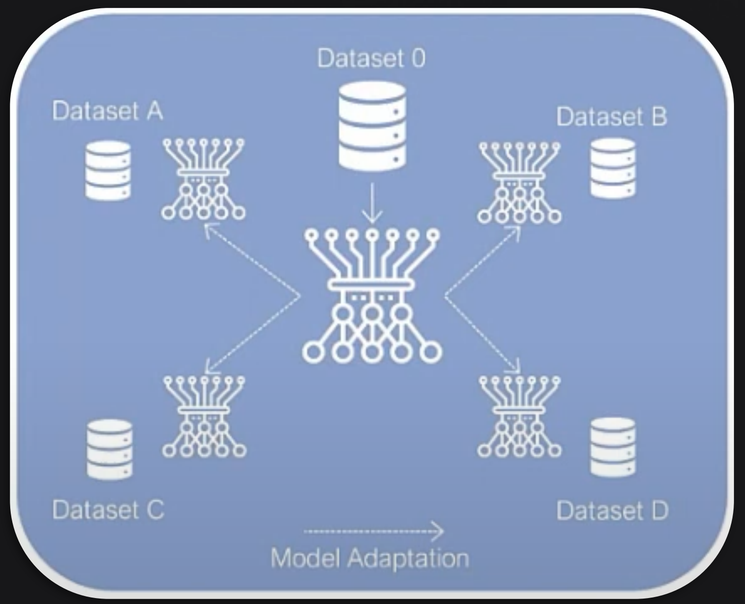
Mapped

USPS -> MNIST

Shaping Datasets – Model Re-purposing

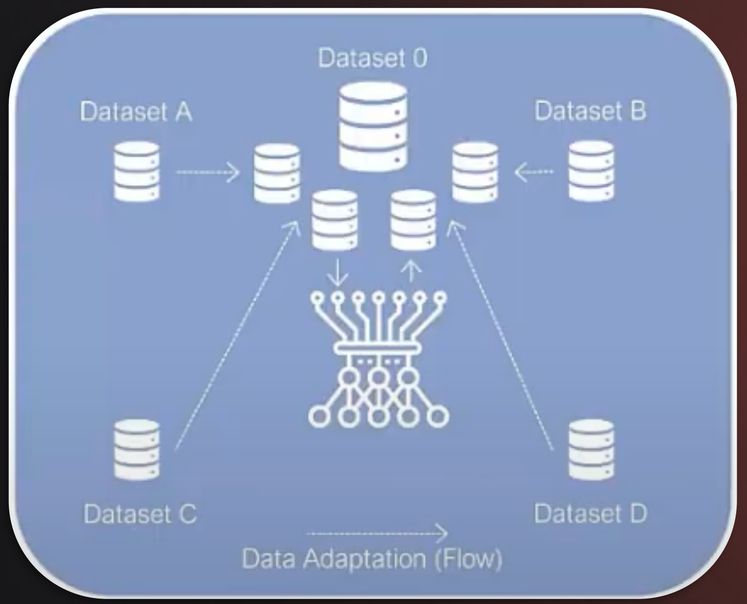
Model-Centric Approach

Clone and adapt



Data-Centric AI

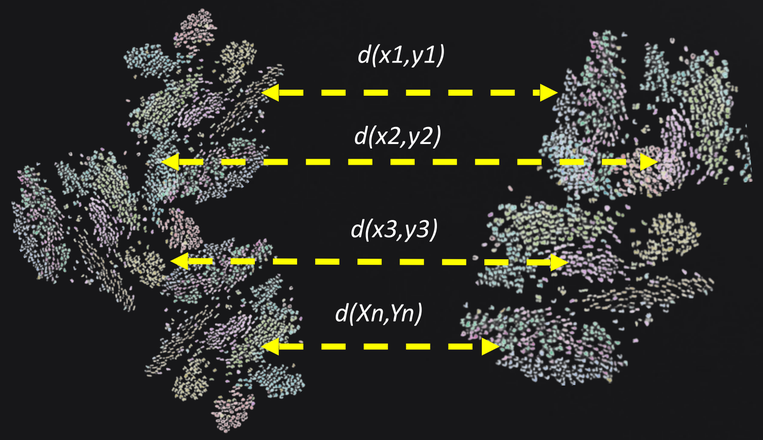
"One model to rule them all"



	Dataset	Original	Flowed
	MNIST	99.5%	-
	USPS	77.1%	99.1%
	KMNIST	5.05%	99.2%
	FMNIST	8.22%	99.1%
	EMNIST	3.62%	36.7%

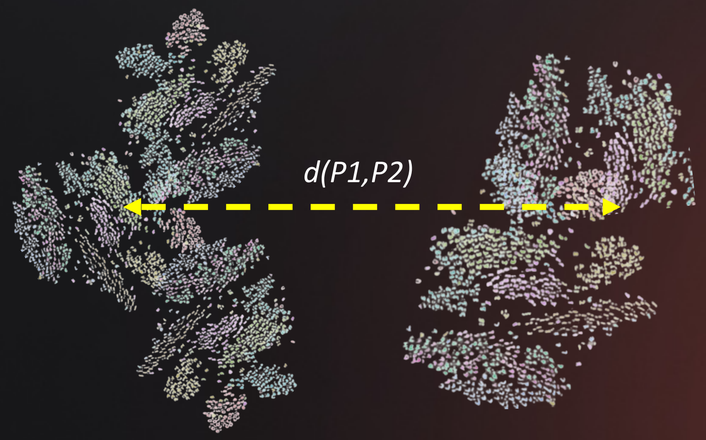
Recap + Takeaways – Optimal Transport

Datasets Correspondence / Alignment



Align unlabeled embedded datasets w/o prior known correspondence

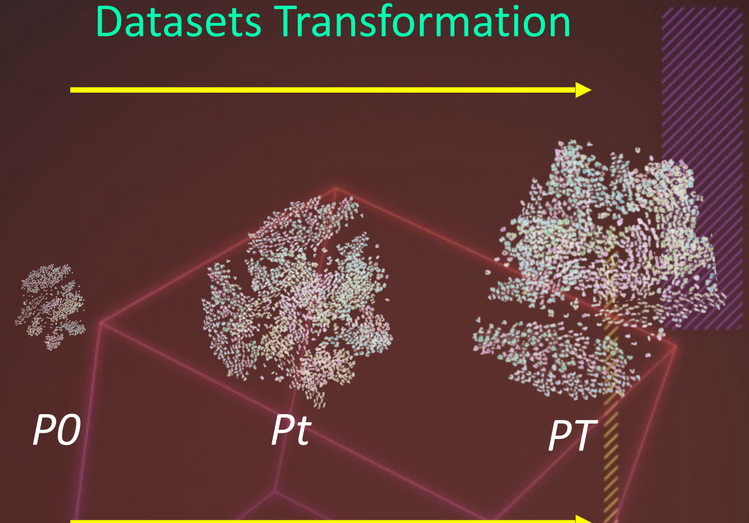
Datasets Distance



OTDD compare labeled datasets even if labels are different

Can be used to guide transfer learning and augmentation processes

Datasets Transformation



A general framework for principled dataset optimization

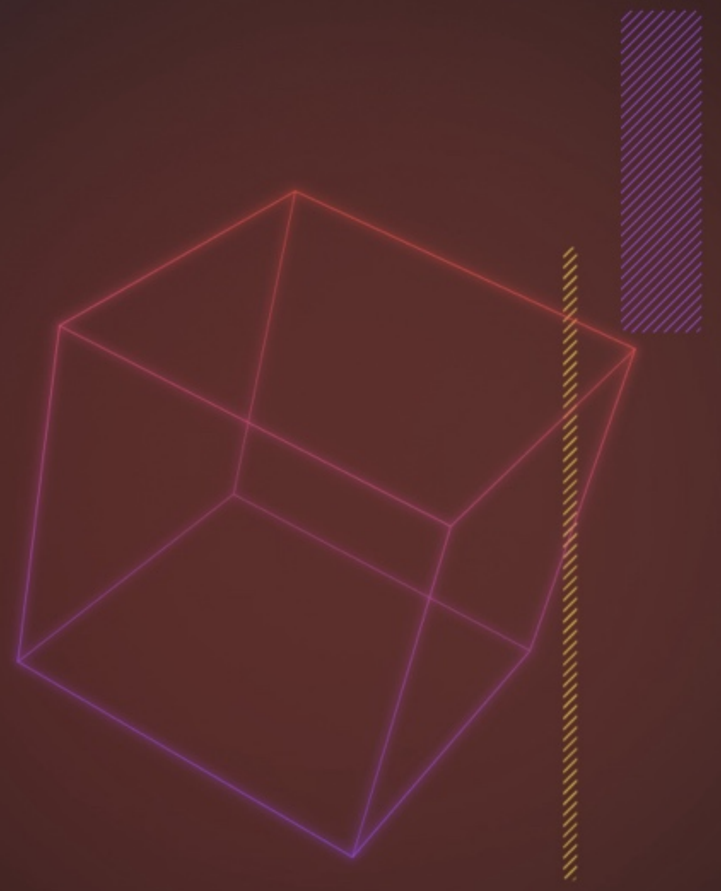
OPTIMAL TRANSPORT

Recap + Takeaways

Code: *github.com/microsoft/otdd*

Thanks to *Alvarez-Melis & Fusi (*Microsoft Research)

- *Geometric Datasets Distances via Optimal Transport, 2020*
- *Gradient Flows in Datasets Space, Arxiv, 2020*
- *Gradient Flows between datasets, ICML, 2021*

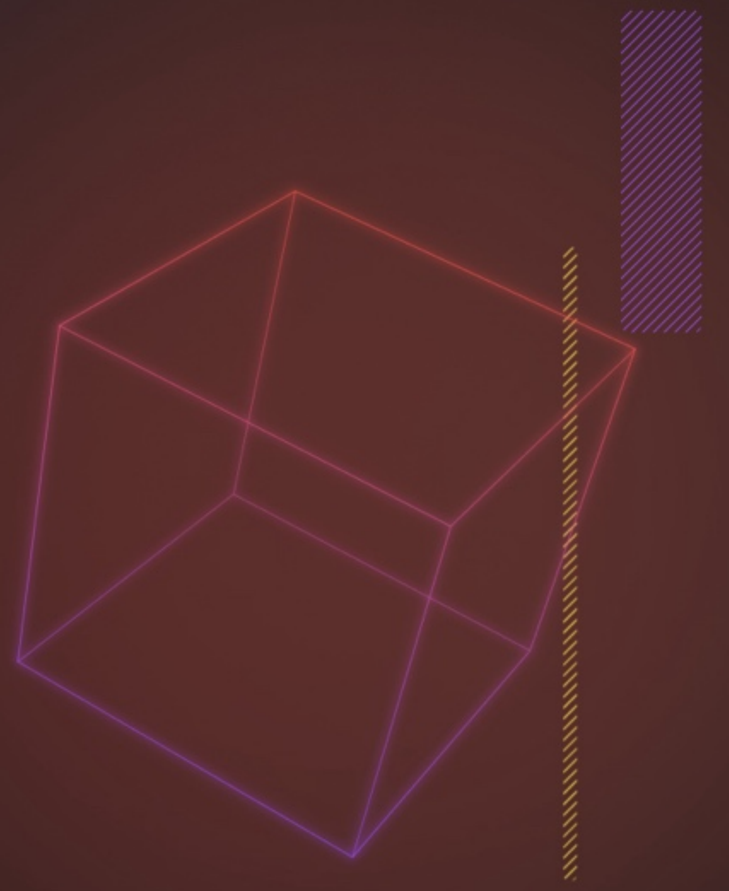


DotNet2022

TECH CONFERENCE

#DotNet2022

Questions & Answers

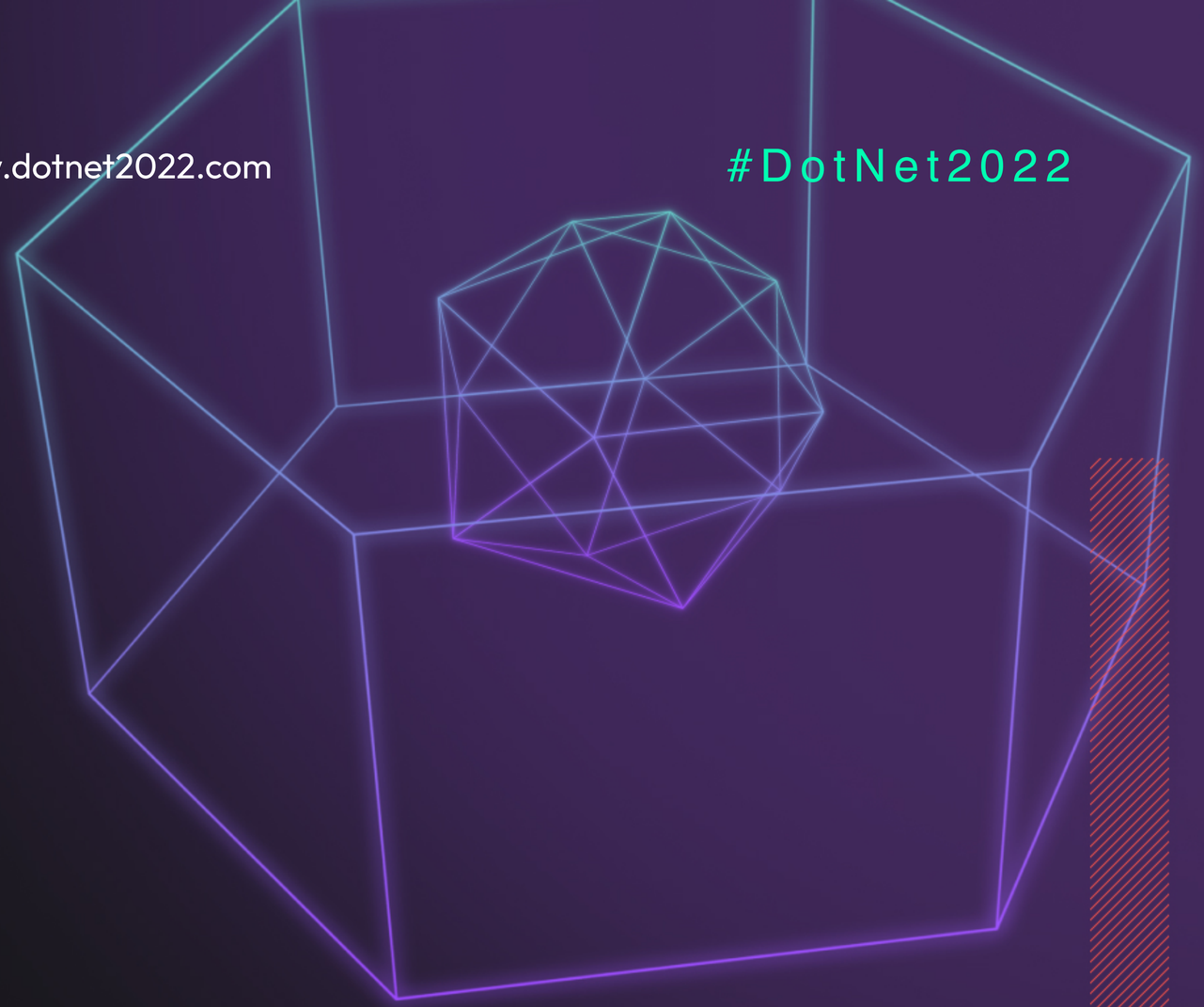


DotNet 2022

TECH CONFERENCE

www.dotnet2022.com

#DotNet2022



Thanks and ...
See you soon!

Thanks also to the sponsors. Without whom this would not have been posible.

plain concepts

Microsoft

intel

LEMON CODE

intelequia

Devs DNA