GDumb: A Simple Approach that Questions Our Progress in Continual Learning



Ameya Prabhu[1] Philip Torr[1] Puneet Dokania[1,2]

University of Oxford[1] & Five AI Ltd.[2]

ECCV'20 ONLINE
23-28 AUGUST 2020

What is Continual Learning?

Input: Each dataset $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$ of *n* samples

Goal : Learn $f_\theta : \mathbf{x} \rightarrow \mathbf{y}$

$$\min_\theta \mathbb{E}_{\mathbf{x},\mathbf{y} \sim \mathcal{D}} L(f_\theta(\mathbf{x}), \mathbf{y})$$

(Standard) Supervised Classification

What happens when it's given a new dataset $\bar{\mathcal{D}}$(having samples with both old and new labels)?

$$\min_\theta \mathbb{E}_{\mathbf{x},\mathbf{y} \sim \mathcal{D} \cup \bar{\mathcal{D}}} L(f_\theta(\mathbf{x}), \mathbf{y})$$

Combine datasets and repeat the process!

What is Continual Learning?

FIVE AI

UNIVERSITY OF OXFORD

(Standard) Supervised Classification

Input: Each dataset $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$ of *n* samples

Goal : Learn $f_\theta : \mathbf{x} \rightarrow \mathbf{y}$

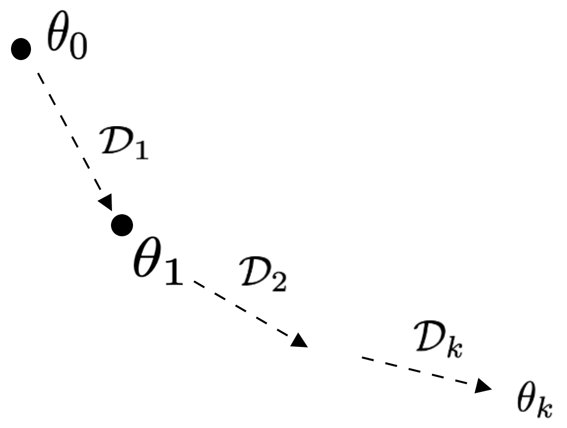
$$\min_\theta \mathbb{E}_{\mathbf{x},\mathbf{y} \sim \cup_{i=1}^{k} \mathcal{D}_i} L(f_\theta(\mathbf{x}), \mathbf{y})$$

It's the same process, repeated *k* times

Continual Classification

Input: A stream of labeled data at each timestep *t*

Goal : Learn $f_\theta : \mathbf{x} \rightarrow \mathbf{y}$



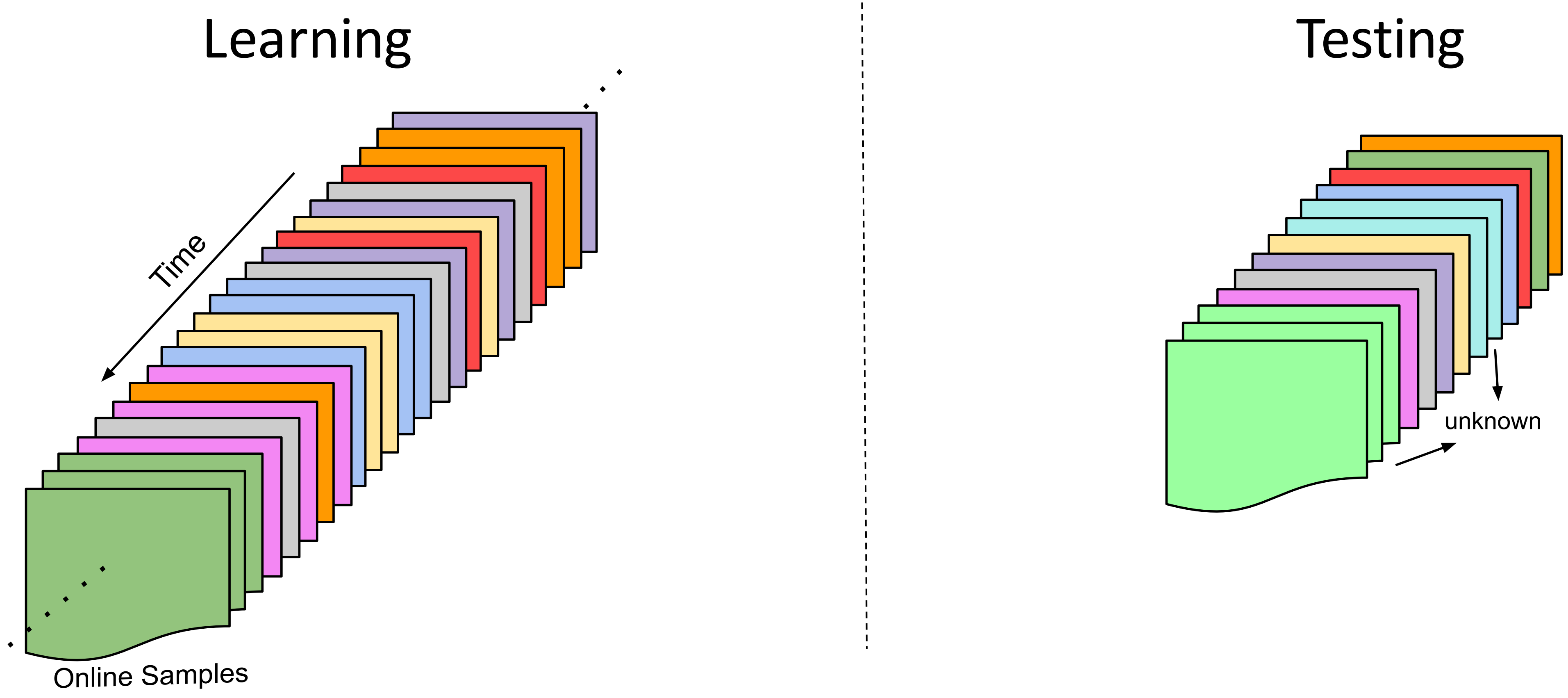
$$\min_\theta \mathbb{E}_{\cup_{i=1}^{k} \mathcal{D}_i} L(f_\theta(\mathbf{x}), \mathbf{y}) \equiv \min_\theta \mathbb{E}_{\mathcal{D}_k} L(f_\theta(\mathbf{x}), \mathbf{y}; \mathbb{K})$$

(Previous knowledge)

Objectives

- Make learning scalable over time
- Mechanisms to add, consolidate & query knowledge ($\mathbb{K}$)

General Continual Learning



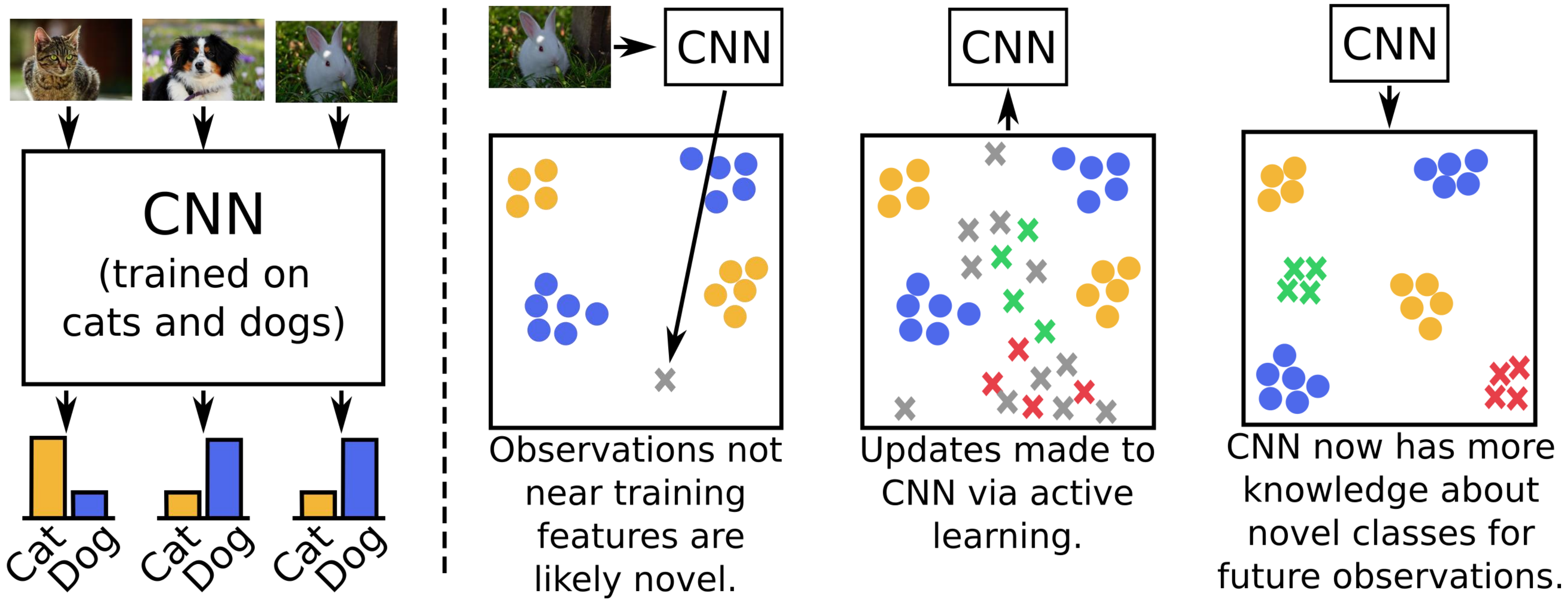
Open-set: The data stream can provide any sample, with any new label, at any time – including at test time

Use-case: Partial information about the classes, consolidate knowledge on-the-fly

General Continual Learning

Supervised Learning

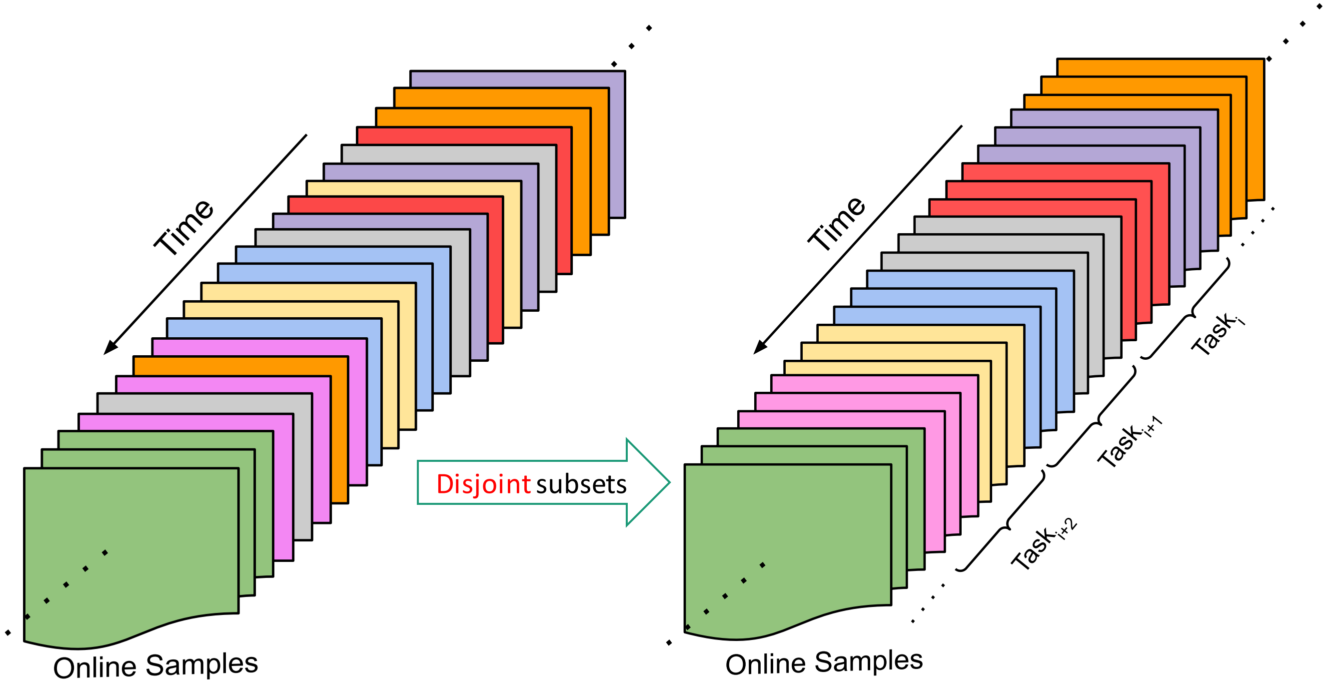
General (Open-Set) Continual Learning



Picture credits: The Importance of Metric Learning for Robotic Vision: Open Set Recognition and Active Learning, ICRA19

Trends in Continual Learning

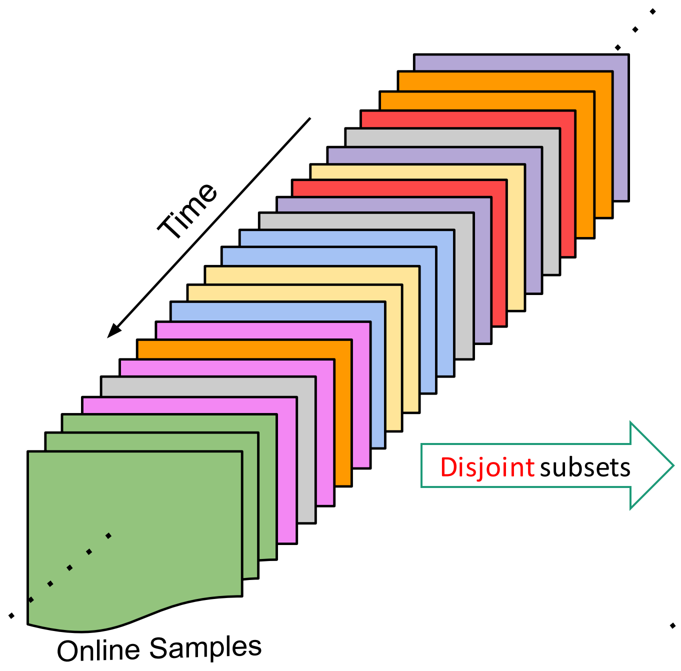
- Classify over all seen labels only (y ∈ Yt)
- Any class (old or new) can come at any time
- Cannot revisit streamed samples again



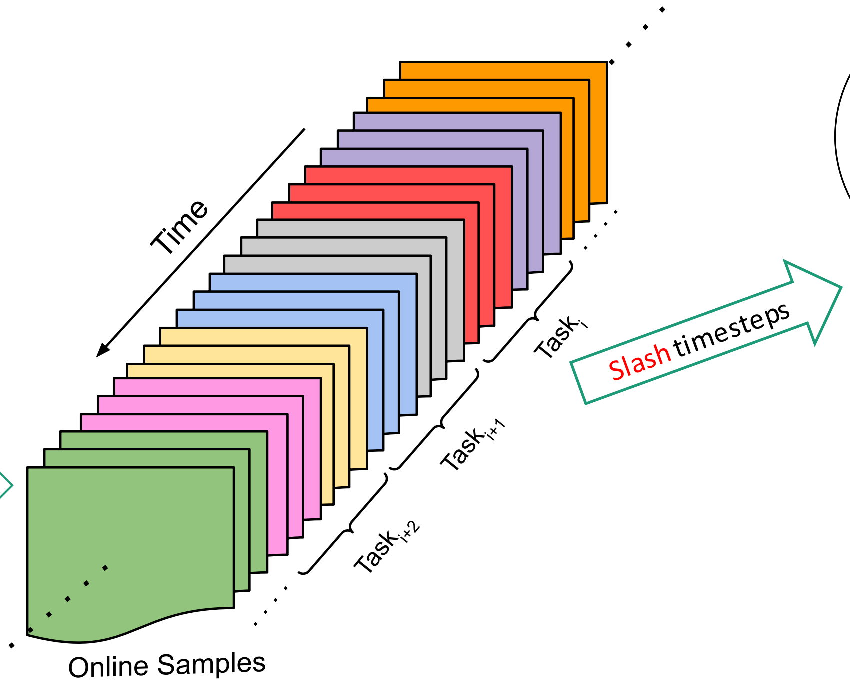
Disjoint Subsets: Clean partitioning into clusters of classes called a task, typically of equal sizes

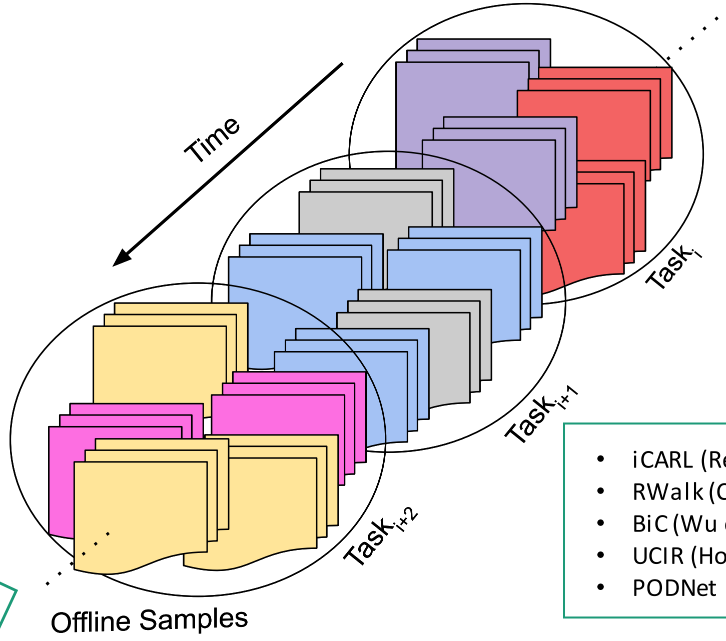
Trends in Continual Learning

- Classify over all seen labels only (y ∈Yt)
- Any class (old or new) can come at any time
- Cannot revisit streamed samples again



- Classify over all seen labels only (y ∈Yt)
- Only new classes can come, with sharp transitions
- Cannot revisit streamed samples again





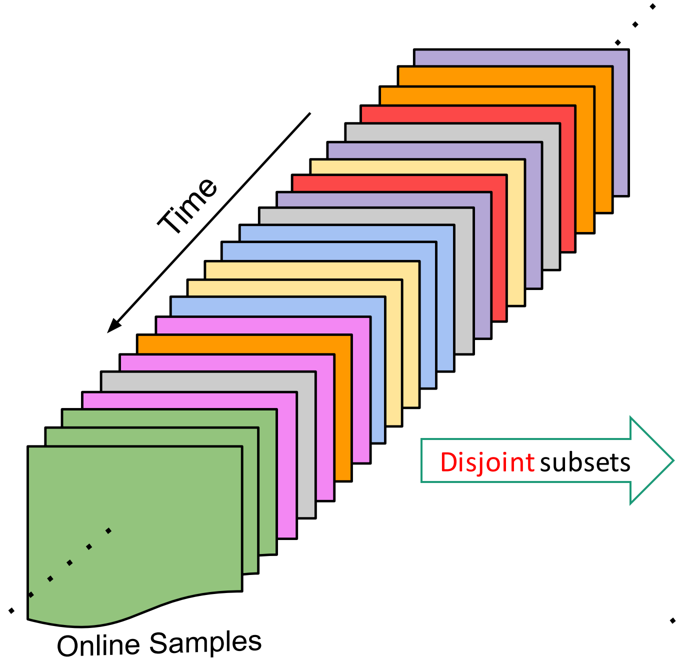
- iCARL (Rebuffi etal., CVPR17)
- RWalk (Chaudhary etal., ECCV18)
- BiC (Wu etal., CVPR19)
- UCIR (Hou etal., CVPR19)
- PODNet (Douillard etal., ECCV20)

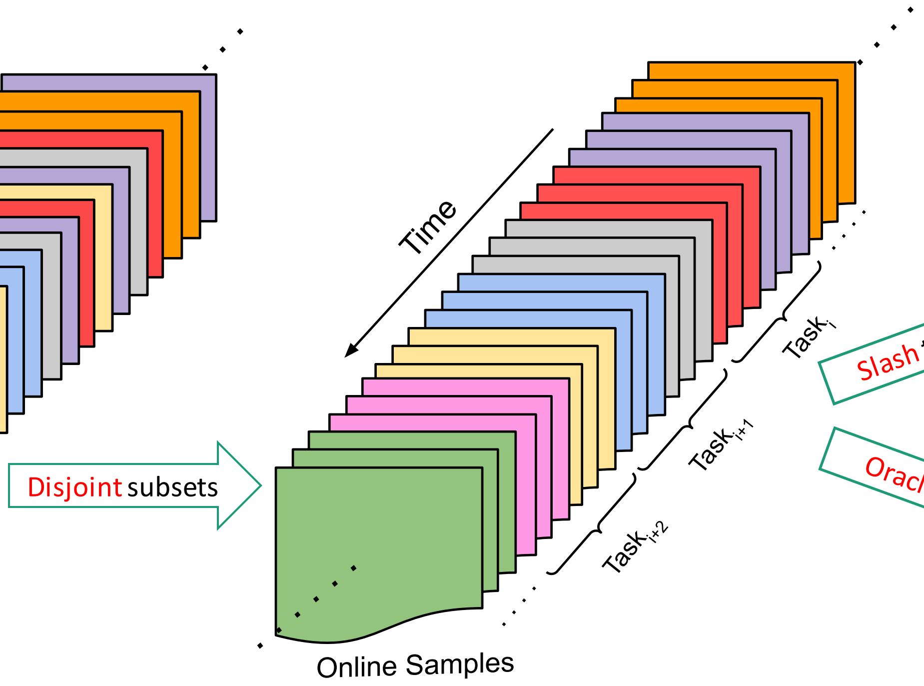
Offline: Clean partitioning into clusters of classes & reduce all timesteps in the same cluster to one

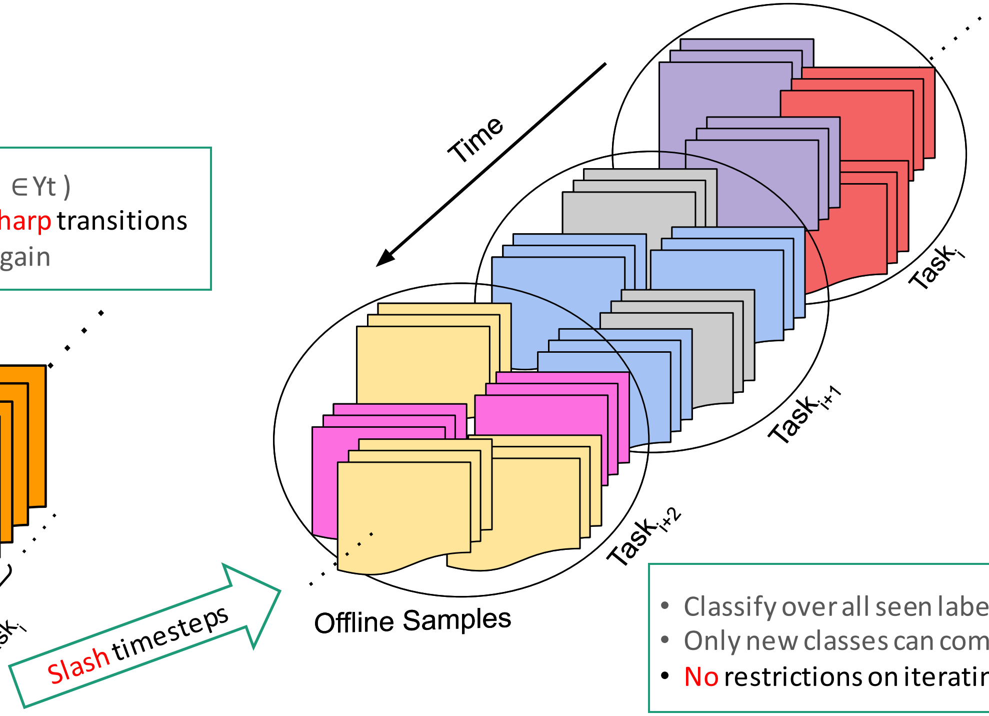
Trends in Continual Learning

- Classify over all seen labels only (y ∈Yt)
- Any class (old or new) can come at any time
- Cannot revisit streamed samples again

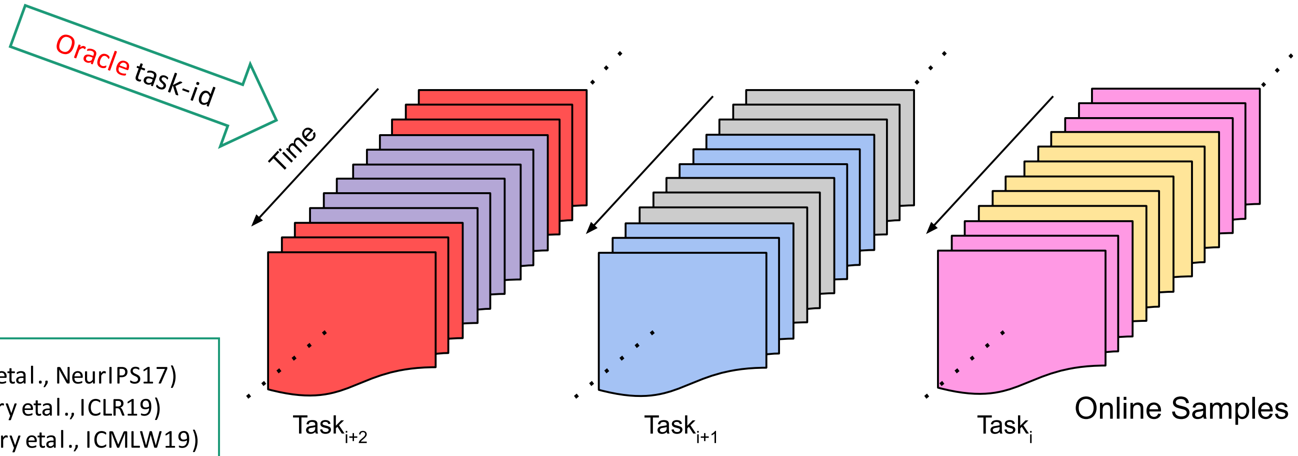
- Classify over all seen labels only (y ∈Yt)
- Only new classes can come, with sharp transitions
- Cannot revisit streamed samples again







- Classify over all seen labels only (y ∈Yt)
- Only new classes can come, with sharp transitions
- No restrictions on iterating over same task samples



- GEM (Lopez-Paz etal., NeurIPS17)
- AGEM (Chaudhary etal., ICLR19)
- TinyER (Chaudhary etal., ICMLW19)

Classifying Literature

Form.	CI-CL	Online	Disjoint	Papers	Regularize	Memory	Distill	Param iso
A	✓	✓	✓	MIR[11], GMED[12]	×	✓	×	×
B	✓	×	✓	LwM[13], DMC[14]	×	×	✓	×
				SDC [15]	✓	×	×	×
				BiC[16], iCARL[4] UCIR[17], EEIL[18] IL2M[19], WA[20] PODNet[21], MCIL[22]	×	✓	✓	×
				RPS-Net[23], iTAML[24]	×	✓	✓	✓
				CGATE[25]	×	✓	×	✓
				RWALK[8]	✓	✓	×	×
C	×	×	✓	PNN[26], DEN[27]	×	×	×	✓
				DGR [28]	×	✓	×	×
				LwF[3]	×	×	✓	×
				P&C[29]	×	×	✓	✓
				APD[30]	✓	×	×	✓
				VCL[31]	✓	✓	×	×
				MAS[32], IMM[33] SI[5], Online-EWC[29] EWC[6]	✓	×	×	×
D	×	✓	✓	TinyER[34], HAL[35]	×	✓	×	×
				GEM[7], AGEM[36]	✓	✓	×	×
E	✓	✓	×	GSS[37]	×	✓	×	×

(Left) Assumptions in formulation

- Disjoint set assumed?
- Task or class-incremental?
- Online or offline?

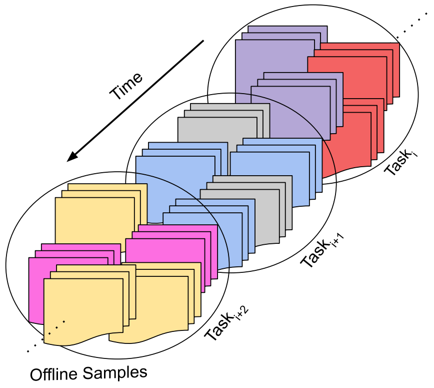
(Right) Strategy to consolidate knowledge

- Regularization?
- Replay?
- Distillation?
- Parameter-isolation?

Classifying Literature

Form.	CI-CL	Online	Disjoint	Papers	Regularize	Memory	Distill	Param iso
A	✓	✓	✓	MIR[11], GMED[12]	×	✓	×	×
				LwM[13], DMC[14]	×	×	✓	×
				SDC [15]	✓	×	×	×
B	✓	×	✓	BiC[16], iCARL[4], UCIR[17], EEIL[18], IL2M[19], WA[20], PODNet[21], MCIL[22]	×	✓	✓	×
				RPS-Net[23], iTAML[24]	×	✓	✓	✓
				CGATE[25]	×	✓	×	✓
				RWALK[8]	✓	✓	×	×
				PNN[26], DEN[27]	×	×	×	✓
				DGR [28]	×	✓	×	×
				LwF[3]	×	×	✓	×
				P&C[29]	×	×	✓	✓
C	×	×	✓	APD[30]	✓	×	×	✓
				VCL[31]	✓	✓	×	×
				MAS[32], IMM[33], SI[5], Online-EWC[29], EWC[6]	✓	×	×	×
D	×	✓	✓	TinyER[34], HAL[35]	×	✓	×	×
				GEM[7], AGEM[36]	✓	✓	×	×
E	✓	✓	×	GSS[37]	×	✓	×	×

For eg: RWALK belongs to this class



Offline, class-incremental, disjoint

RWALK aims to mitigate forgetting using regularization with help of memory

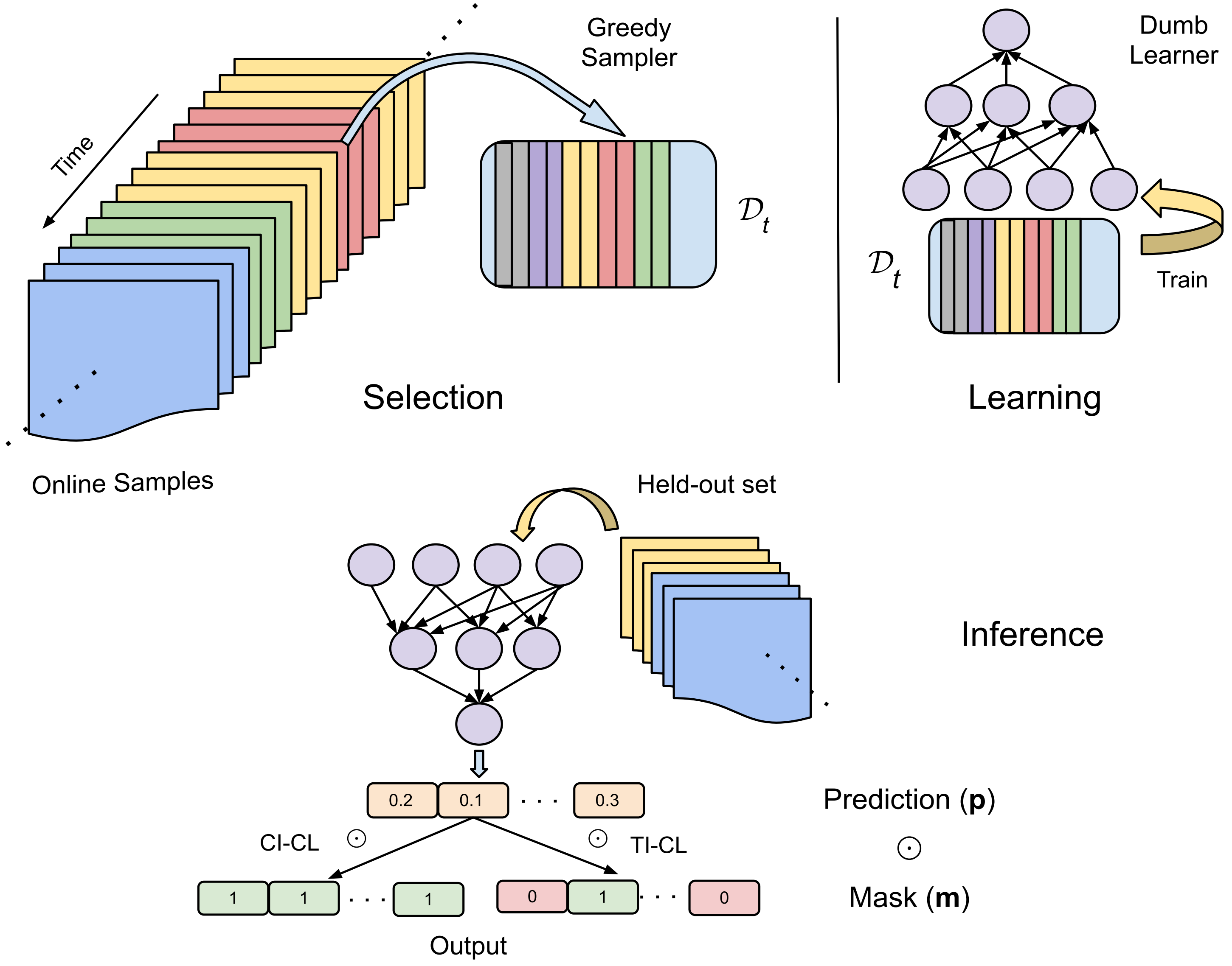
Classifying Literature

Form.	CI-CL	Online	Disjoint	Papers	Regularize	Memory	Distill	Param iso
A	✓	✓	✓	MIR[11], GMED[12]	×	✓	×	×
B	✓	×	✓	LwM[13], DMC[14]	×	×	✓	×
				SDC [15]	✓	×	×	×
				BiC[16], iCARL[4], UCIR[17], EEIL[18], IL2M[19], WA[20], PODNet[21], MCIL[22]	×	✓	✓	×
				RPS-Net[23], iTAML[24]	×	✓	✓	✓
				CGATE[25]	×	✓	×	✓
				RWALK[8]	✓	✓	×	×
C	×	×	✓	PNN[26], DEN[27]	×	×	×	✓
				DGR [28]	×	✓	×	×
				LwF[3]	×	×	✓	×
				P&C[29]	×	×	✓	✓
				APD[30]	✓	×	×	✓
				VCL[31]	✓	✓	×	×
				MAS[32], IMM[33], SI[5], Online-EWC[29], EWC[6]	✓	×	×	×
D	×	✓	✓	TinyER[34], HAL[35]	×	✓	×	×
				GEM[7], AGEM[36]	✓	✓	×	×
E	✓	✓	×	GSS[37]	×	✓	×	×

Typical CL Algorithms

- Evaluated on one specific formulation
  - Formulation oversimplified & restricted
  - Algorithms often fail to generalize
  - Are the scenarios practical?
- *Very* sensitive to hyperparameters
  - Can't tweak when deployed
- *Very* computationally intensive
  - Why not train a supervised model directly?

GDumb: A Simple, Unifying Approach



GDumb

Greedy Balancing Sampler

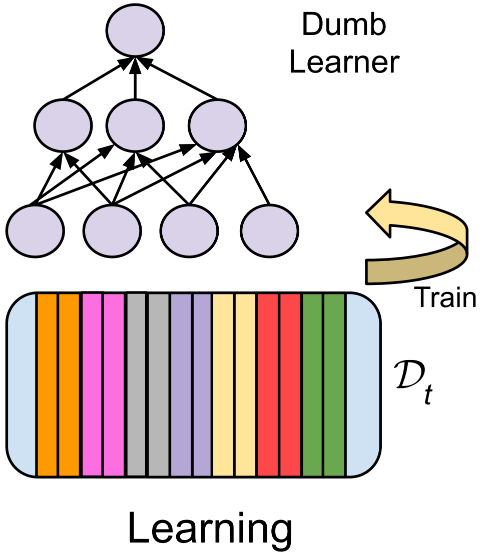
- Greedily stores samples in memory
- Balances #samples across classes

Dumb Learner

- When asked, trains a model *from scratch* only using current memory samples
- Combines predictions with oracle task-information via a binary mask at inference

Greedy Sampler & Dumb Learner

- GDumb has no explicit model designed for:
  - *Nothing* to reduce forgetting
  - *Nothing* to improve intransigence
- Same, simple learning
  - *No* task-incremental training
  - *No* offline training
  - *No* disjoint sampling
- No hyperparameter tuning!



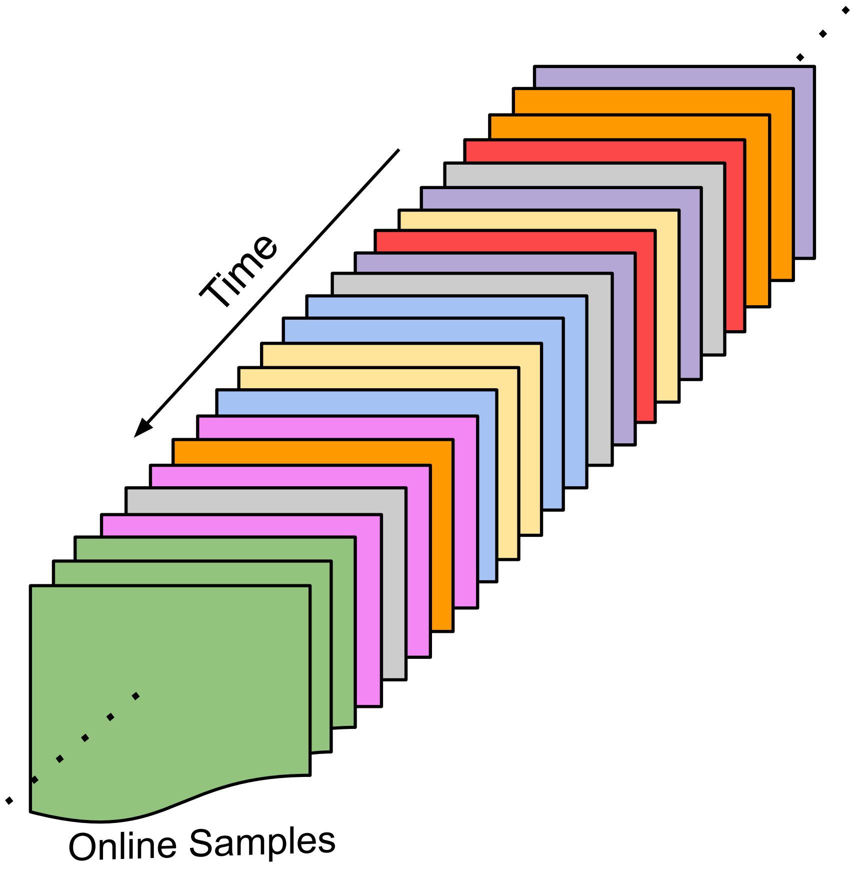
Experimental Setup

Form.	Designed in	Model (Dataset)	memory (k)	Metric	CI-CL	Online	Disjoint
A1	MIR	MLP-400 (MNIST); ResNet18 (CIFAR10)	300, 500; 200, 500, 1000	Acc. (at end)			
A2	GMED	MLP-400 (MNIST); ResNet18 (CIFAR10)	500; 500	Acc. (at end)	✓	✓	✓
A3	ARM	MLP-400 (MNIST); ResNet18 (CIFAR10)	500; 1000	Acc. (at end)			
B1	Hsu etal. RPS-Net	MLP-400 (MNIST); ResNet18 (SVHN)	4400	Acc. (at end)			
B2	iCARL	ResNet32 (CIFAR100)	2000	Acc. (avg in t)	✓	×	✓
B3	PODNet	ResNet32 (CIFAR100); ResNet18 (ImageNet100)	1000-2000 (+20) x50	Acc. (avg in t)			
C1	Hsu etal.	MLP-400 (MNIST)	4400	Acc. (at end)	×	×	✓
C2	CSDF	Many (TinyImageNet)	4500,9000	Acc. (at end)			
D	AGEM	ResNet-18-S (CIFAR10)	0-1105 (+65) x17	Acc. (at end)	×	✓	✓
E	GSS	MLP-100 (MNIST); ResNet-18 (CIFAR10)	300; 500	Acc. (at end)	✓	✓	×

Evaluate on 10 popular, diverse formulations

- Same network & memory
- No hyperparameter tuning
  - SGD
  - lr: 5e-2 → 5e-4
  - SGDR schedular
  - Decay: 1e-6
  - Batch size: 16
- No formulation restrictions used for training

Minimal Assumptions: Comparisons

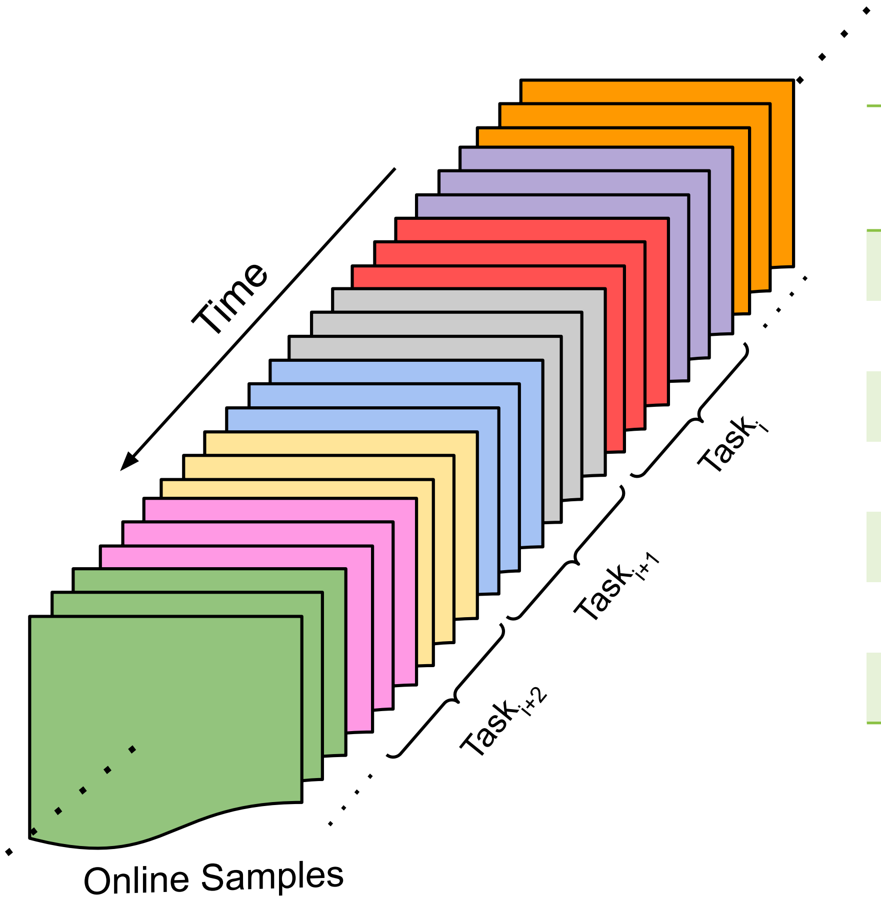


- GSS (Aljundi etal., NeurIPS19)

Method	MNIST	CIFAR10
Reservoir	69.1	-
GSS-Clust	-	25.0
FSS-Clust	-	26.0
GSS-IQP	76.5	29.6
GSS-Greedy	78.0	29.6
GDumb	**88.9**	**45.8**
(+Increase)	**(+10.9)**	**(+16.2)**

Beats best competitor by 10-15% points

+Disjoint Sets Assumption



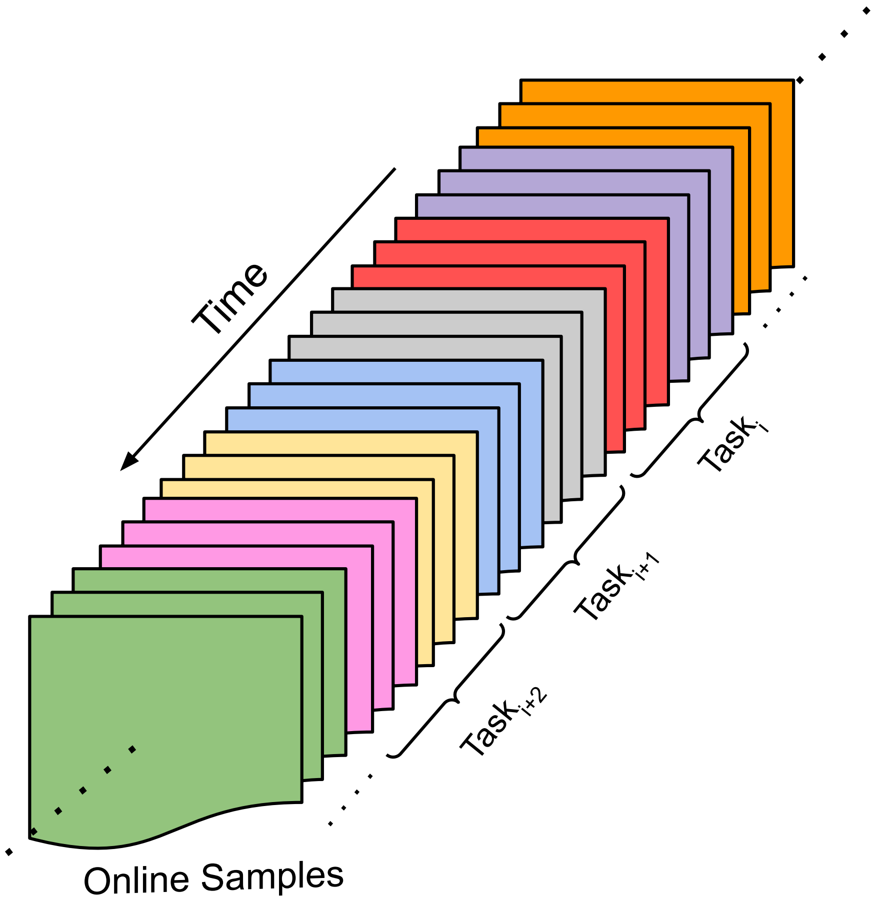
- MIR (Aljundi etal., NeurIPS19)

Method (k)	MNIST (500)
GEN	75.5 ± 1.3
GEN-MIR	81.6 ± 0.9
ER	82.1 ± 2.4
GEM	86.3 ± 1.8
ER-MIR	87.6 ± 0.7
GDumb	91.9 ± 0.5
(+Increase)	(+4.3)

Method (k)	CIFAR10 (200)	(500)	(1000)
GEM	16.8 ± 1.1	17.1 ± 1.0	17.5 ± 1.6
iCARL	28.6 ± 1.2	33.7 ± 1.6	32.4 ± 2.1
ER	27.5 ± 1.2	33.1 ± 1.7	41.3 ± 1.9
ER-MIR	29.8 ± 1.1	40.0 ± 1.1	47.6 ± 1.1
ER5	-	-	42.4 ± 1.1
ER-MIR5	-	-	49.3 ± 0.1
GDumb	**35.0 ± 0.6**	**45.8 ± 0.9**	**61.3 ± 1.7**
(+Increase)	(+5.2)	(+5.8)	(+11.0)

Beats previous best which uses disjoint set assumption by 4-11% points (lower margin)

+Disjoint Sets Assumption

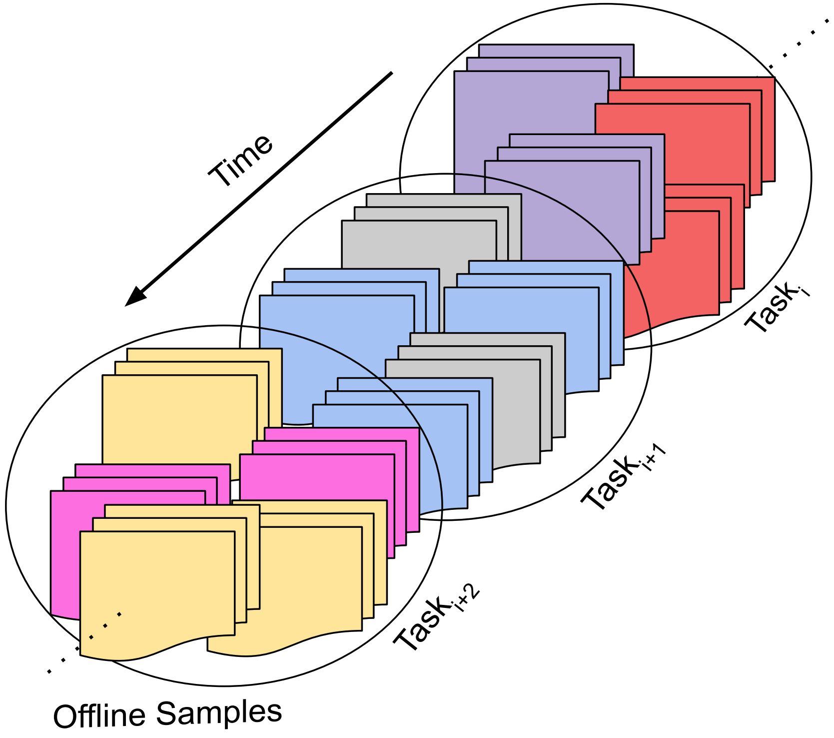


- GMED (Jin et al., Arxiv, July20)

Method (k)	MNIST (500)	CIFAR10 (500)
Fine tuning	18.8 ± 0.6	18.5 ± 0.2
AGEM	29.0 ± 5.3	18.5 ± 0.6
BGD	13.5 ± 5.1	18.2 ± 0.5
GEM	87.2 ± 1.3	20.1 ± 1.4
GSS-Greedy	84.2 ± 2.6	28.0 ± 1.3
HAL	77.9 ± 4.2	32.1 ± 1.5
ER	81.0 ± 2.3	33.3 ± 1.5
MIR	84.9 ± 1.7	34.5 ± 2.0
GMED (ER)	82.7 ± 2.1	35.0 ± 1.5
GMED (MIR)	87.9 ± 1.1	35.5 ± 1.9
GDumb	**91.9 ± 0.5**	**45.8 ± 0.9**
(+Increase)	(+4.0)	(+10.3)

Beats parallel work which uses disjoint assumption by 4-10% points

+Disjoint, Offline Sets Assumption

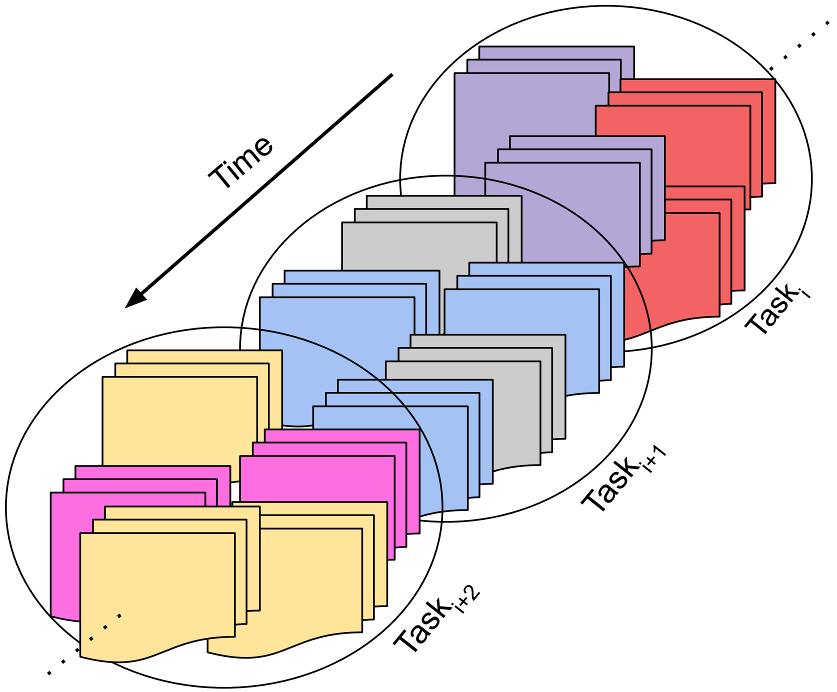


- Hsu etal., NeurIPS18 CL-W)

Method	MNIST	SVHN
MAS	19.5 ± 0.3	17.3
SI	19.7 ± 0.1	17.3
EWC	19.8 ± 0.1	18.2
Online-EWC	19.8 ± 0.04	18.5
LwF	24.2 ± 0.3	-
DGR	91.2 ± 0.3	-
DGR+Distill	91.8 ± 0.3	-
GEM	92.2 ± 0.1	75.6
RtF	92.6 ± 0.2	-
RPS-Net	96.2	88.9
OvA-INN	96.4	-
iTAML	97.9	94.0
GDumb	97.8 ± 0.2	93.4 ± 0.4

Beats all competitors inspite disjoint & offline assumptions, matching iTAML performance

+Disjoint, Offline Sets Assumption



iCARL (Rebuffi etal., CVPR17) PODNet (Douillard etal., ECCV20)

Method/CIFAR100	**10 tasks, 10 cls**	**50 tasks, 1 cls**
DMC++	56.8 ± 0.9	-
iCARL	58.8 ± 1.9	44.2 ± 1.0
WA	62.6	-
EEIL	63.4 ± 1.6	-
BiC	63.8	47.1 ± 1.5
UCIR (CNN)	-	49.3 ± 0.3
PODNet (CNN)	-	58.0 ± 0.5
GDumb	45.2 ± 1.7	58.4 ± 0.8
(Diff w) iCARL, BiC	**-13.6**, **-18.6**	**+14.2**, **+11.3**

+30!

When tasks were 10, we were ~15-20% lower When tasks increase to 50, we perform 10-15% higher
Illustrates: BiC/iCARL don't work beyond formulations having less timesteps (tasks)

Questioning Progress in Continual Learning

Method	MNIST	
k	(300)	(500)
MLP-100		
FSS-Clust [37]	75.8 ± 1.7	83.4 ± 2.6
GSS-Clust [37]	75.7 ± 2.2	83.9 ± 1.6
GSS-IQP [37]	75.9 ± 2.5	84.1 ± 2.4
GSS-Greedy [37]	82.6 ± 2.9	84.8 ± 1.8
GDumb (Ours)	**88.9 ± 0.6**	**90.0 ± 0.4**
MLP-400		
GEN [43]	-	75.5 ± 1.3
GEN-MIR [11]	-	81.6 ± 0.9
ER [44]	-	82.1 ± 1.5
GEM [7]	-	86.3 ± 1.4
ER-MIR [11]	-	87.6 ± 0.7
GDumb (Ours)	-	**91.9 ± 0.5**

Method	MNIST
(k)	(4400)
GEM [7]	98.42 ± 0.10
EWC [6]	98.64 ± 0.22
SI [5]	99.09 ± 0.15
Online EWC [29]	99.12 ± 0.11
MAS [32]	99.22 ± 0.21
DGR [28]	99.50 ± 0.03
LwF [3]	99.60 ± 0.03
DGR+Distil [28]	99.61 ± 0.02
RtF	99.66 ± 0.03
GDumb	**99.77 ± 0.03**

(C1)

Method	Parameters	Regularization	Accuracy
No stored samples			
mean-IMM [33]	3.5M	none	32.42
mode-IMM [33]	9.0M	dropout	42.41
SI [5]	3.5M/9.0M	L2/dropout	43.74
HAT [51]	3.5M/9.0M	L2	44.19
EWC [6]	613K	none	45.13
LwF [3]	9.0M	L2	48.11
EBLL [52]	9.0M	L2	48.17
MAS [32]	3.5M/9.0M	none	48.98
PackNet [53]	613K/3.5M	L2/dropout	55.96
k=4500			
GEM [7]	613K/3.5M	none/dropout	44.23
GDumb	834K	cutmix	45.50
iCARL [4]	613K/3.5M	dropout	48.55
k=9000			
GEM [7]	613K/3.5M	none/dropout	45.27
iCARL [4]	613K/3.5M	dropout	49.94
GDumb	834K	cutmix	**57.27**

Method	CIFAR10		
k	(200)	(500)	(1000)
GEM [7]	16.8 ± 1.1	17.1 ± 1.0	17.5 ± 1.6
iCARL [4]	28.6 ± 1.2	33.7 ± 1.6	32.4 ± 2.1
ER [44]	27.5 ± 1.2	33.1 ± 1.7	41.3 ± 1.9
ER-MIR [11]	29.8 ± 1.1	40.0 ± 1.1	47.6 ± 1.1
ER5 [11]	-	-	42.4 ± 1.1
ER-MIR5 [11]	-	-	49.3 ± 0.1
GDumb (Ours)	**35.0 ± 0.6**	**45.8 ± 0.9**	**61.3 ± 1.7**

(A1)

Method	CIFAR100
(k)	(1105)
RWalk [8]	40.9 ± 3.9
EWC [6]	42.4 ± 3.0
Base	42.9 ± 2.0
MAS [32]	44.2 ± 2.3
SI [5]	47.1 ± 4.4
iCARL [4]	50.1
S-GEM [36]	56.2
PNN [26]	59.2 ± 0.8
GEM [7]	61.2 ± 0.7
A-GEM [36]	63.1 ± 1.2
TinyER [34]	68.5 ± 0.6
GDumb	60.3 ± 0.85

(D)

Method	MNIST	CIFAR10
Reservoir [43]	69.12	-
GSS-Clust [37]	-	25.0
FSS-Clust [37]	-	26.0
GSS-IQP [37]	76.49	29.6
GSS-Greedy [37]	77.96	29.6
GDumb (Ours)	**88.93**	**45.8**

(E)

Method	MNIST	CIFAR-10
k	(500)	(500)
Fine tuning	18.8 ± 0.6	18.5 ± 0.2
AGEM [36]	29.0 ± 5.3	18.5 ± 0.6
BGD [48]	13.5 ± 5.1	18.2 ± 0.5
GEM [7]	87.2 ± 1.3	20.1 ± 1.4
GSS-Greedy [37]	84.2 ± 2.6	28.0 ± 1.3
HAL [35]	77.9 ± 4.2	32.1 ± 1.5
ER [44]	81.0 ± 2.3	33.3 ± 1.5
MIR [11]	84.9 ± 1.7	34.5 ± 2.0
GMED (ER) [12]	82.7 ± 2.1	35.0 ± 1.5
GMED (MIR) [12]	87.9 ± 1.1	35.5 ± 1.9
GDumb (Ours)	**91.9 ± 0.5**	**45.8 ± 0.9**

(A2)

Method	MNIST		CIFAR10	
	Memory	Accuracy	Memory	Accuracy
Finetune	0	18.8 ± 0.5	0	15.0 ± 3.1
GEN [28]	4.58	79.3 ± 0.6	34.5	15.3 ± 0.5
GEN-MIR [11]	4.31	82.1 ± 0.3	38.0	15.3 ± 1.2
LwF [3]	1.91	33.3 ± 2.5	4.38	19.2 ± 0.3
ADI [47]	1.91	55.4 ± 2.6	4.38	24.8 ± 0.9
ARM [41]	1.91	56.2 ± 3.5	4.38	26.4 ± 1.2
ER [44]	0.39	83.2 ± 1.9	3.07	41.3 ± 1.9
ER-MIR [11]	0.39	85.6 ± 2.0	3.07	47.6 ± 1.1
iCarl [4] (5 iter)	-	-	3.07	32.4 ± 2.1
GEM [7]	0.39	86.3 ± 0.1	3.07	17.5 ± 1.6
GDumb (ours)	**0.39**	**91.9 ± 0.5**	**3.07**	**61.3 ± 1.7**

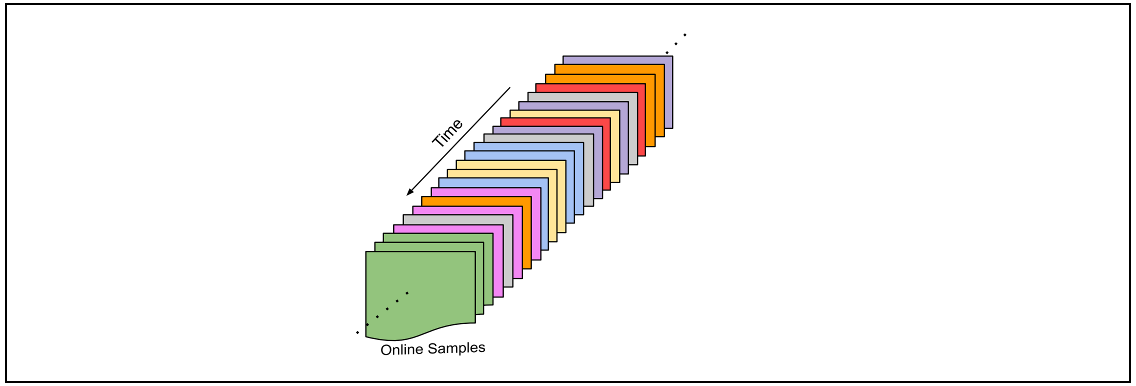
(A3)

Possible failure modes:

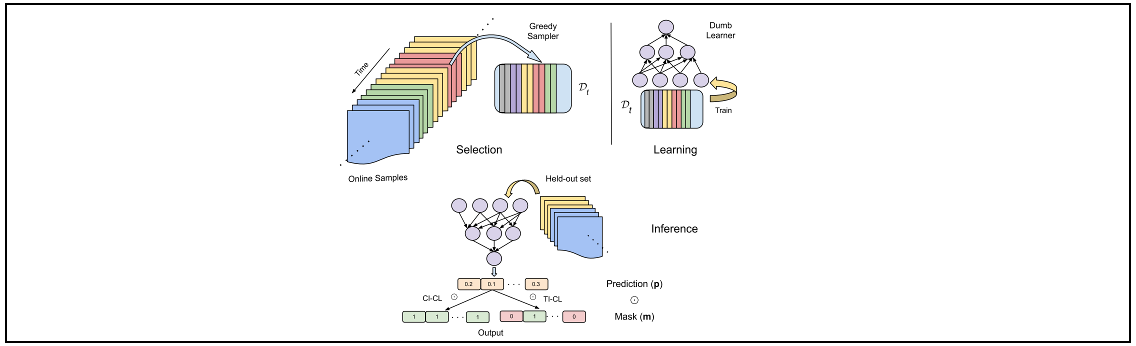
- Bad evaluation (metrics, ..) ?
- Too simplistic/restrictive formulations?
- Heavily tailored approaches?

It's alarming that simple GDumb outperforms tailored algorithms on formulations they were designed for!

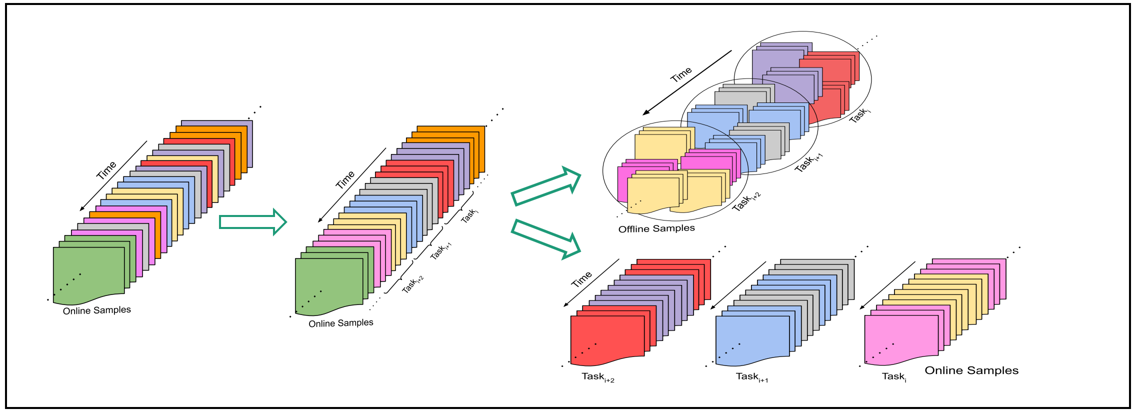
Summary: Our Contributions



A General CL Formulation



GDumb: A Simple, Unifying Approach



Quirks & Assumptions of Recent Formulations

Thank You!

Questions?