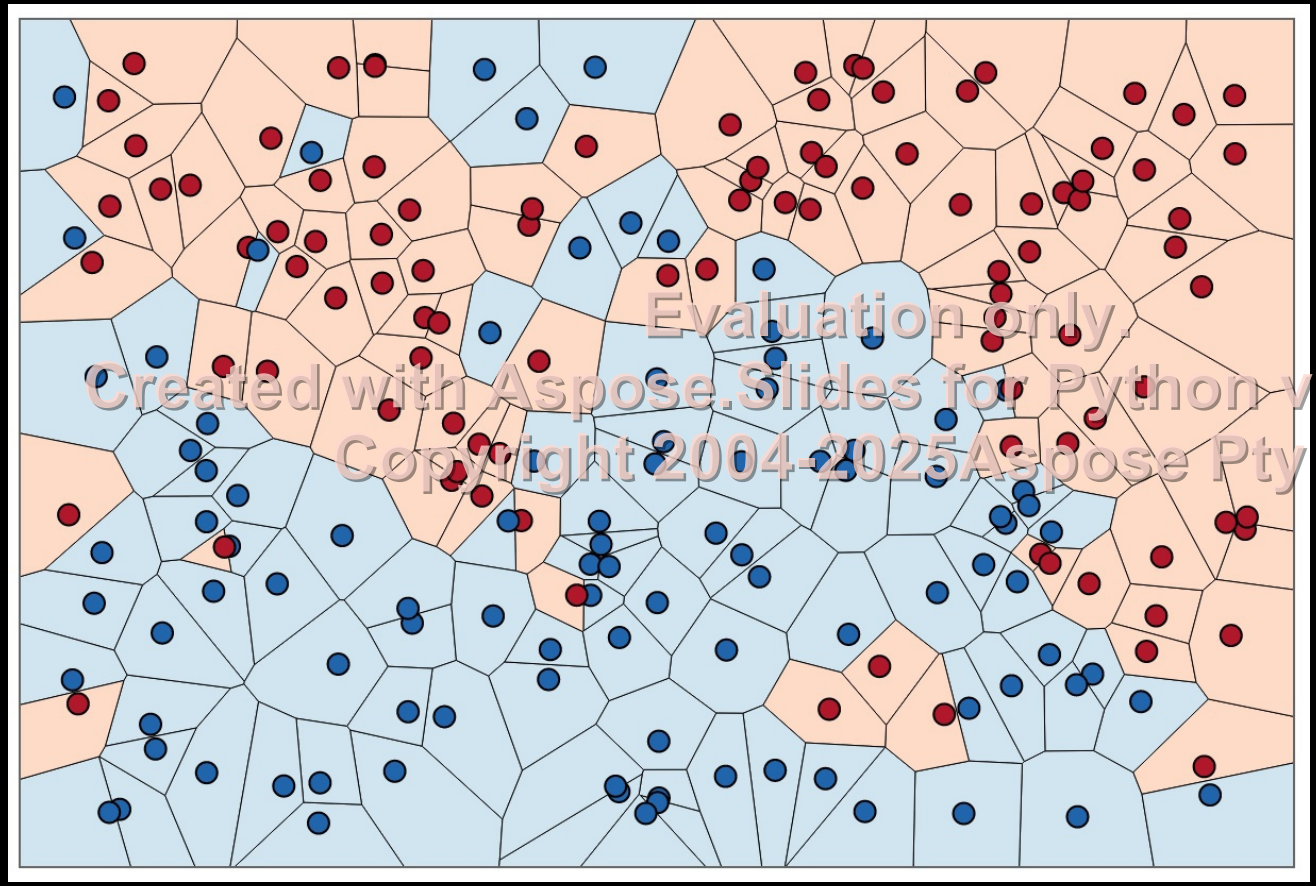
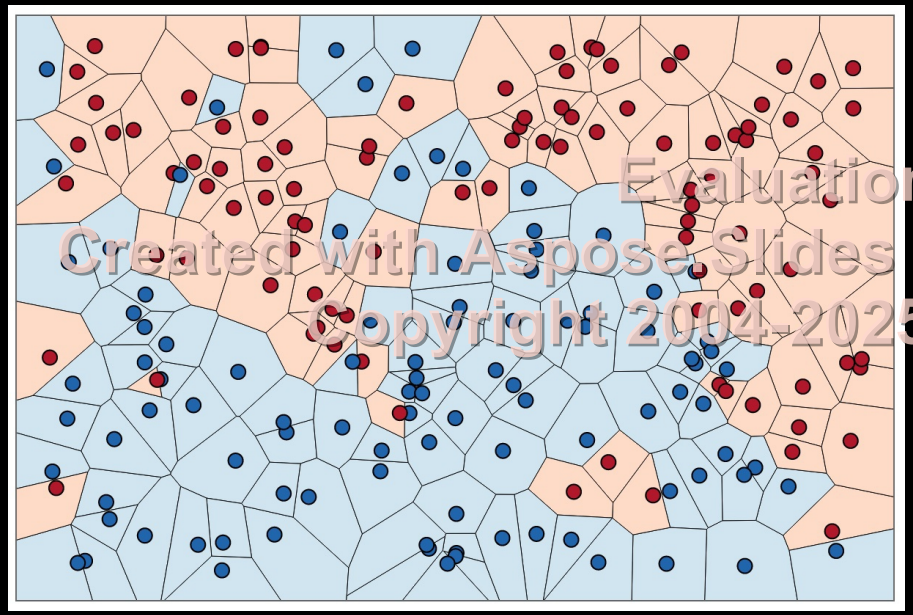
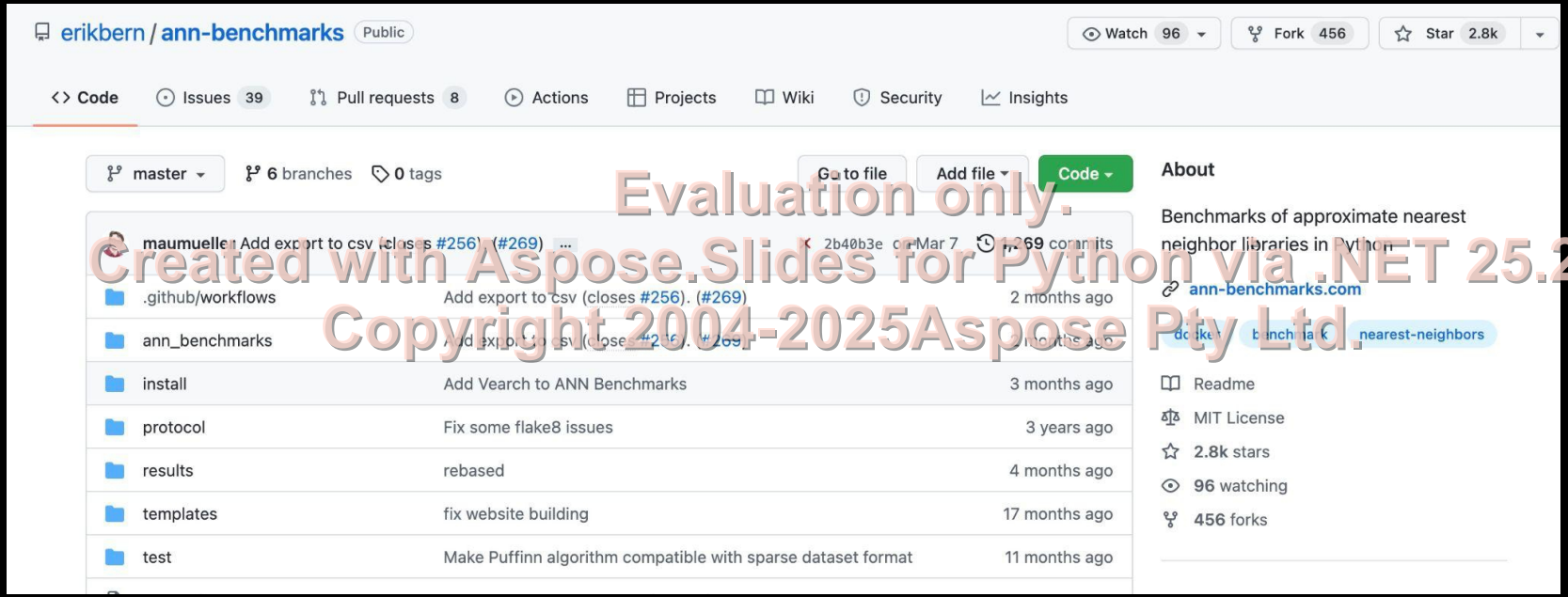


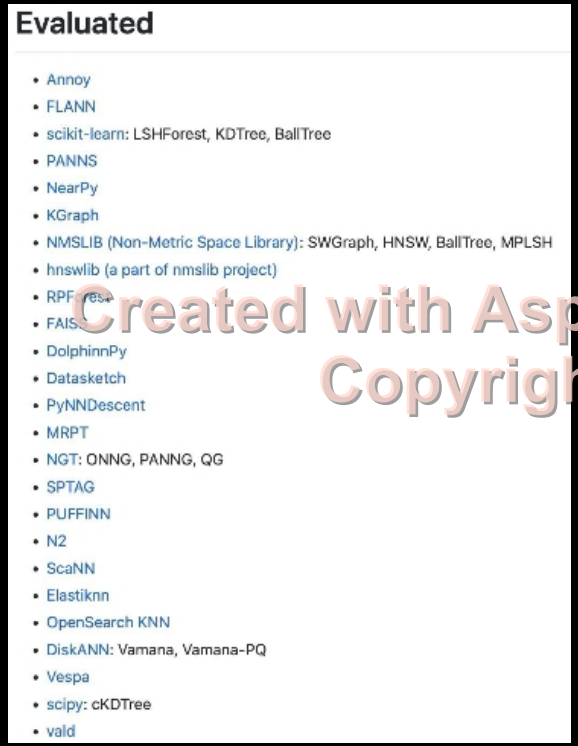


SMILE IDENTITY

SMILE IDENTITY



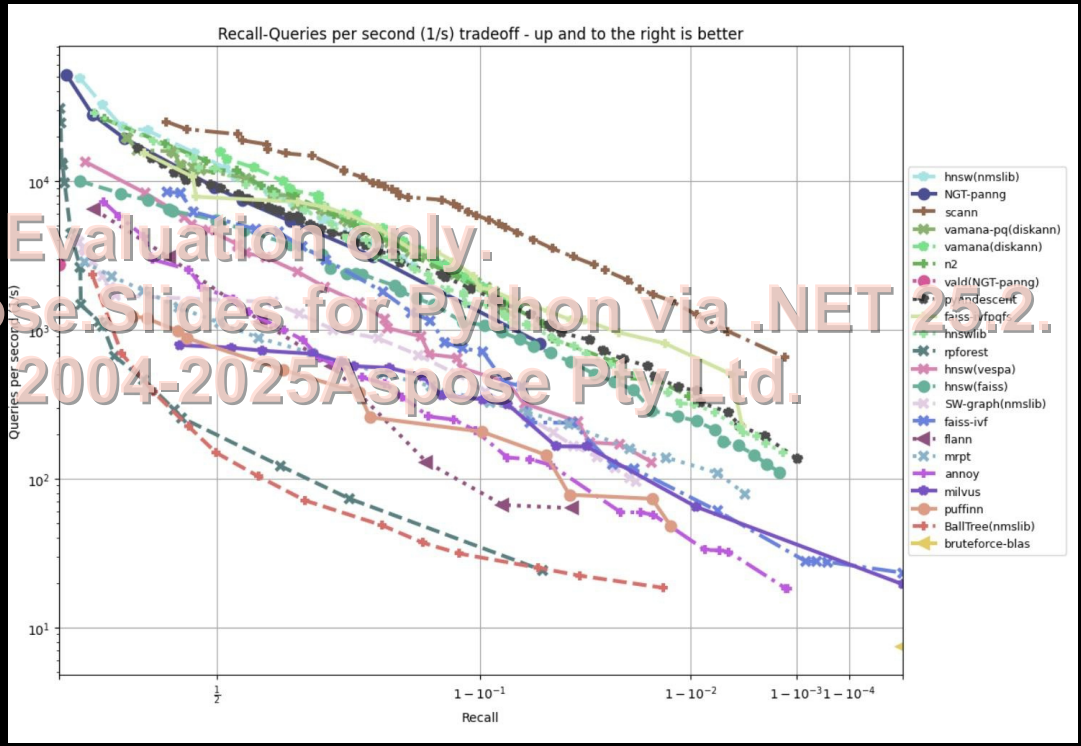
http://github.com/erikbern/ann-benchmarks



http://github.com/erikbern/ann-benchmarks

Evaluated

- Annoy
- FLANN
- scikit-learn: LSHForest, KDTree, BallTree
- PANNS
- NearPy
- KGraph
- NMSLIB (Non-Metric Space Library): SWGraph, HNSW, BallTree, MPLSH
- hnswlib (a part of nmslib project)
- RPF[illegible]
- FAIS[illegible]
- DolphinnPy
- Datasketch
- PyNNDescent
- MRPT
- NGT: ONNG, PANNG, QG
- SPTAG
- PUFFINN
- N2
- ScaNN
- Elastiknn
- OpenSearch KNN
- DiskANN: Vamana, Vamana-PQ
- Vespa
- scipy: cKDTree
- vald



http://github.com/erikbern/ann-benchmarks

SMILE IDENTITY

Data sets

We have a number of precomputed data sets for this. All data sets are pre-split into train/test and come with ground truth data in the form of the top 100 neighbors. We store them in a HDF5 format:

Dataset	Dimensions	Train size	Test size	Neighbors	Distance	Download
DEEP1B	96	9,990,000	10,000	100	Angular	HDF5 (3.6GB)
Fashion-MNIST	784	60,000	10,000	100	Euclidean	HDF5 (217MB)
GIST	960	1,000,000	1,000	100	Euclidean	HDF5 (3.6GB)
GloVe	25	1,183,514	10,000	100	Angular	HDF5 (121MB)
GloVe	50	1,183,514	10,000	100	Angular	HDF5 (235MB)
GloVe	100	1,183,514	10,000	100	Angular	HDF5 (463MB)
GloVe	200	1,183,514	10,000	100	Angular	HDF5 (918MB)
Kosarak	27983	74,962	500	100	Jaccard	HDF5 (2.0GB)
MNIST	784	60,000	10,000	100	Euclidean	HDF5 (217MB)
NYTimes	256	290,000	10,000	100	Angular	HDF5 (301MB)
SIFT	128	1,000,000	10,000	100	Euclidean	HDF5 (501MB)
Last.fm	65	292,385	50,000	100	Angular	HDF5 (135MB)

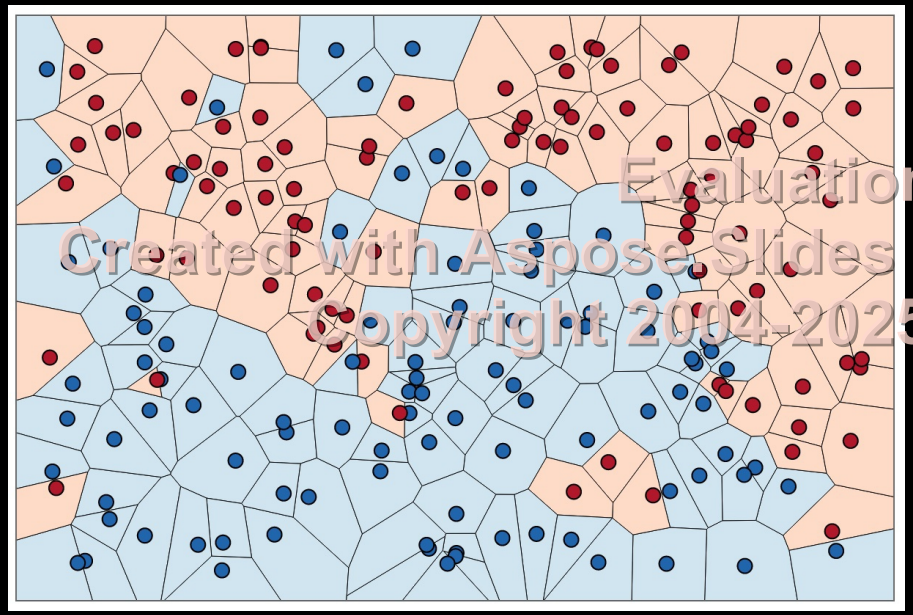
http://github.com/erikbern/ann-benchmarks

Data sets

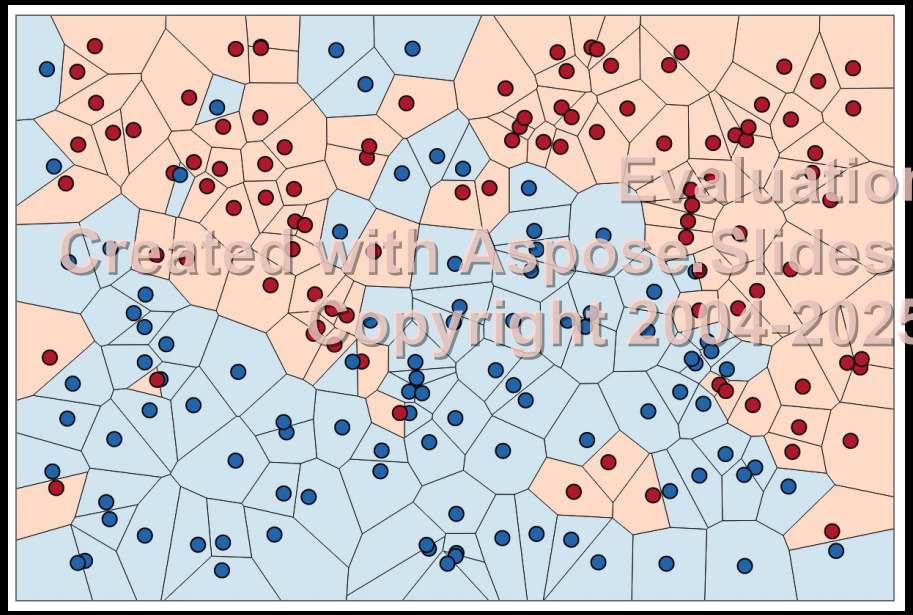
We have a number of precomputed data sets for this. All data sets are pre-split into train/test and come with ground truth data in the form of the top 100 neighbors. We store them in a HDF5 format:

Dataset	Dimensions	Train size	Test size	Neighbors	Distance	Download
DEEP1B	96	9,990,000	10,000	100	Angular	HDF5 (3.6GB)
Fashion-MNIST	784	60,000	10,000	100	Euclidean	HDF5 (217MB)
GIST	960	1,000,000	1,000	100	Euclidean	HDF5 (3.6GB)
GloVe	25	1,183,514	10,000	100	Angular	HDF5 (121MB)
GloVe	50	1,183,514	10,000	100	Angular	HDF5 (235MB)
GloVe	100	1,183,514	10,000	100	Angular	HDF5 (463MB)
GloVe	200	1,183,514	10,000	100	Angular	HDF5 (918MB)
Kosarak	27983	74,962	500	100	Jaccard	HDF5 (2.0GB)
MNIST	784	60,000	10,000	100	Euclidean	HDF5 (217MB)
NYTimes	256	290,000	10,000	100	Angular	HDF5 (301MB)
SIFT	128	1,000,000	10,000	100	Euclidean	HDF5 (501MB)
Last.fm	65	292,385	50,000	100	Angular	HDF5 (135MB)

http://github.com/erikbern/ann-benchmarks

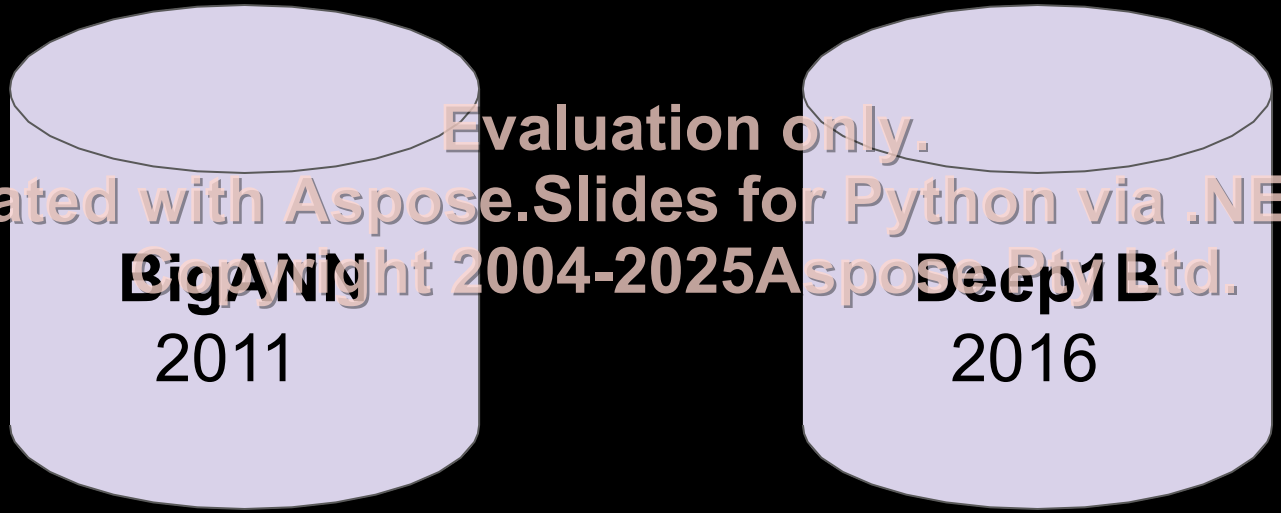


SMILE IDENTITY



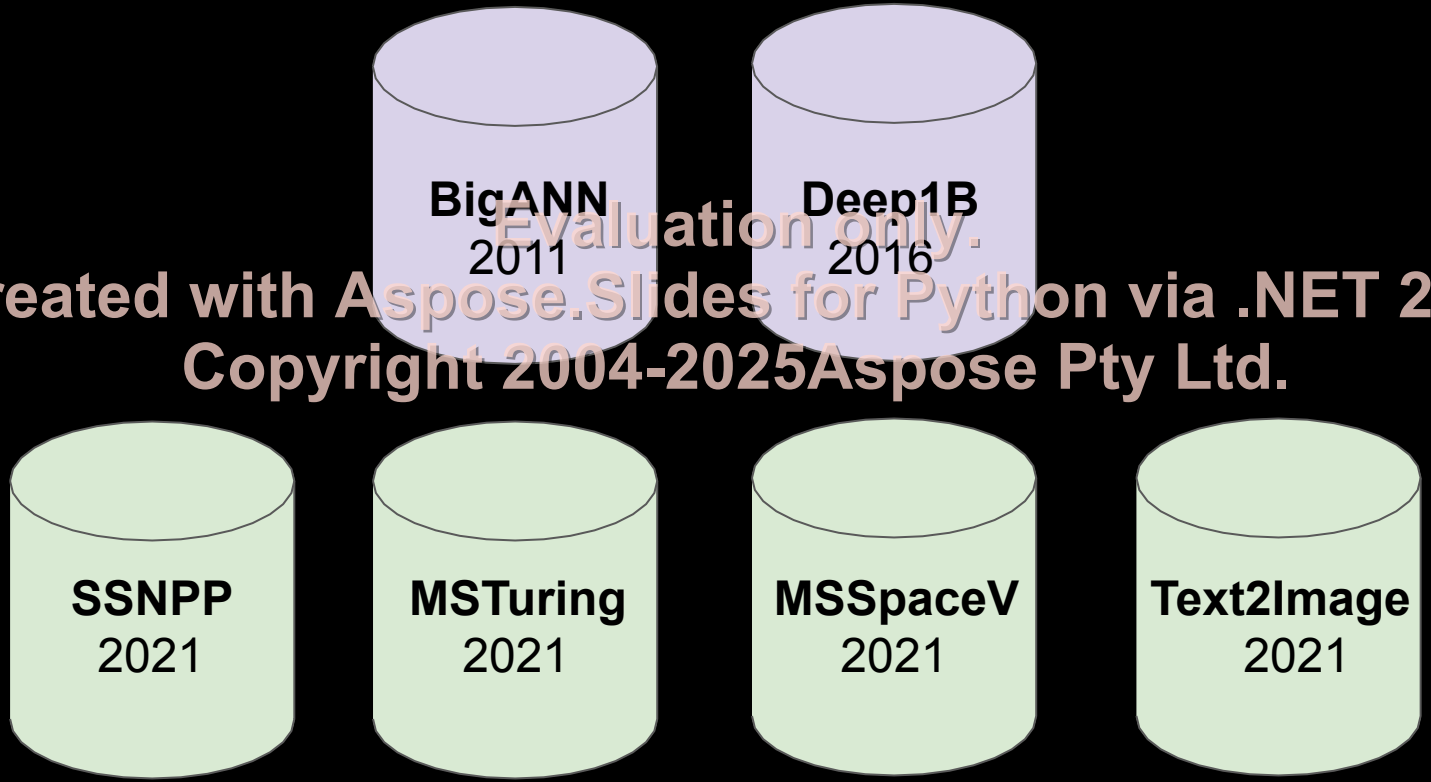
SMILE IDENTITY

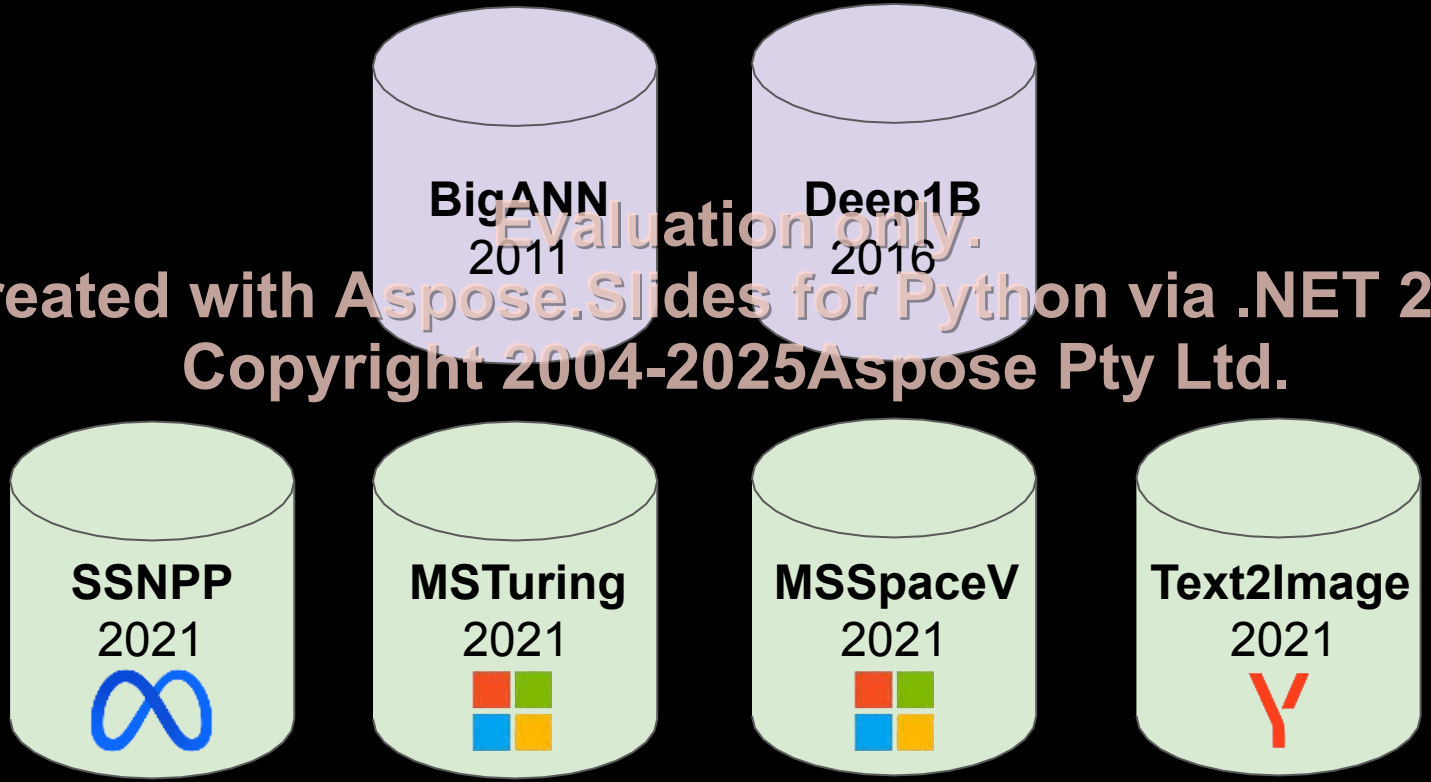
SMILE IDENTITY



Jegou, Tavenard,
Douze, Laurent (INRIA)

Babenko, Lempitsky (CVPR)





SMILE IDENTITY

Dataset	Type	d	Dist.	Query Type
BIGANN	uint8	128	L2	k-NN
Facebook SimSearchNet++	uint8	256	L2	Range
Microsoft Turing ANNS	float32	100	L2	k-NN
Microsoft SpaceV	int8	100	L2	k-NN
DEEP	float32	96	L2	k-NN
Yandex Text-to-Image	float32	200	IP	k-NN

SMILE IDENTITY

Track #1

Datacenter Server

64 GB RAM

Track #2

Track #3

SMILE IDENTITY

Track #1

Datacenter Server

64 GB RAM

Track #2

Datacenter Server

64 GB RAM

2TB HTD SSD

Track #3

SMILE IDENTITY

Track #1

Datacenter Server

64 GB RAM

Track #2

Datacenter Server

64 GB RAM

2TB HTD SSD

Track #3

Any Hardware

Evidence of Cost

Power Monitoring

SMILE IDENTITY

Track #1

Datacenter Server

64 GB RAM

Track #2

Datacenter Server

64 GB RAM

2TB HTD SSD

Track #3

Any Hardware

Evidence of Cost

Power Monitoring

- Limited Hardware Resource Budget
- Ranked By ***Recall***

Track #1

Datacenter Server

64 GB RAM

Track #2

Datacenter Server

64 GB RAM

2TB HTD SSD

Track #3

Any Hardware

Evidence of Cost

Power Monitoring

- Limited Hardware Resource Budget
- Ranked By ***Recall***

- ***Recall***, QPS, Power, Cost

SMILE IDENTITY

Size	vCPU's	Memory: GiB	Temp storage (SSD) GiB	Max data disks	Max cached and temp storage throughput: IOPS/MBps (cache size in GiB)	Max uncached disk throughput: IOPS/MBps	Max burst uncached disk throughput IOPS/MBps
Standard_F2s_v2[1]	2	4	16	4	4000/31 (32)	3200/47	4000/200
Standard_F4s_v2	4	8	32	8	8000/63 (64)	6400/95	8000/200
Standard_F8s_v2	8	16	64	16	16000/127 (128)	12800/190	16000/400
Standard_F16s_v2	16	32	128	32	32000/255 (256)	25600/380	32000/800
Standard_F32s_v2	32	64	256	32	64000/512 (512)	51200/750	64000/1600
Standard_F48s_v2	48	96	384	32	96000/768 (768)	76800/1100	80000/2000
Standard_F64s_v2	64	128	512	32	128000/1024 (1024)	80000/1100	80000/2000
Standard_F72s_v2[2, 3]	72	144	576	32	144000/1152 (1520)	80000/1100	80000/2000

FSv2 Compute Optimized Series

SMILE IDENTITY

Microsoft Azure

Size	vCPU's	Memory: GiB	Temp storage (SSD) GiB	Max data disks	Max cached and temp storage throughput: IOPS/MBps (cache size in GiB)	Max uncached disk throughput: IOPS/MBps	Max burst uncached disk throughput IOPS/MBps
[illegible]	2	4	16	4	4000/31 (32)	3200/47	4000/200
[illegible]	4	8	[illegible]	8	[illegible]	[illegible]	[illegible]
Standard_F8s_v2	8	16	64	16	16000/127 (128)	12800/190	16000/400
Standard_F16s_v2	16	32	128	32	32000/255 (256)	25600/380	32000/800
Standard_F32s_v2	32	64	256	32	64000/512 (512)	51200/750	64000/1600
Standard_F48s_v2	48	96	384	32	96000/768 (768)	76800/1100	80000/2000
Standard_F64s_v2	64	128	512	32	128000/1024 (1024)	80000/1100	80000/2000
Standard_F72s_v2[2, 3]	72	144	576	32	144000/1152 (1520)	80000/1100	80000/2000

FSv2 Compute Optimized Series

SMILE IDENTITY

Microsoft Azure

Size	vCPU	Memory (GiB)	Temp disk[1] (GiB)	NVMe Disks[2]	NVMe Disk throughput[3] (Read IOPS/MBps)	Uncached data disk throughput (IOPs/MBps)[4]	Max burst uncached data disk throughput (IOPs/MB
Standard_L8s_v2	8	64	80	1x1.92 TB	400000/2000	8000/160	8000/1280
Standard_L16s_v2	16	128	160	2x1.92 TB	800000/4000	16000/320	16000/128
Standard_L32s_v2	32	256	320	4x1.92 TB	1.5M/8000	32000/640	32000/128
Standard_L48s_v2	48	384	480	6x1.92 TB	2.2M/14000	48000/960	48000/200
Standard_L64s_v2	64	512	640	8x1.92 TB	2.9M/16000	64000/1280	64000/200
Standard_L80s_v2[6]	80	640	800	10x1.92TB	3.8M/20000	80000/1400	80000/200

LSv2 Storage Optimized w/High Throughput NVM SSD

SMILE IDENTITY

SMILE IDENTITY

SMILE IDENTITY

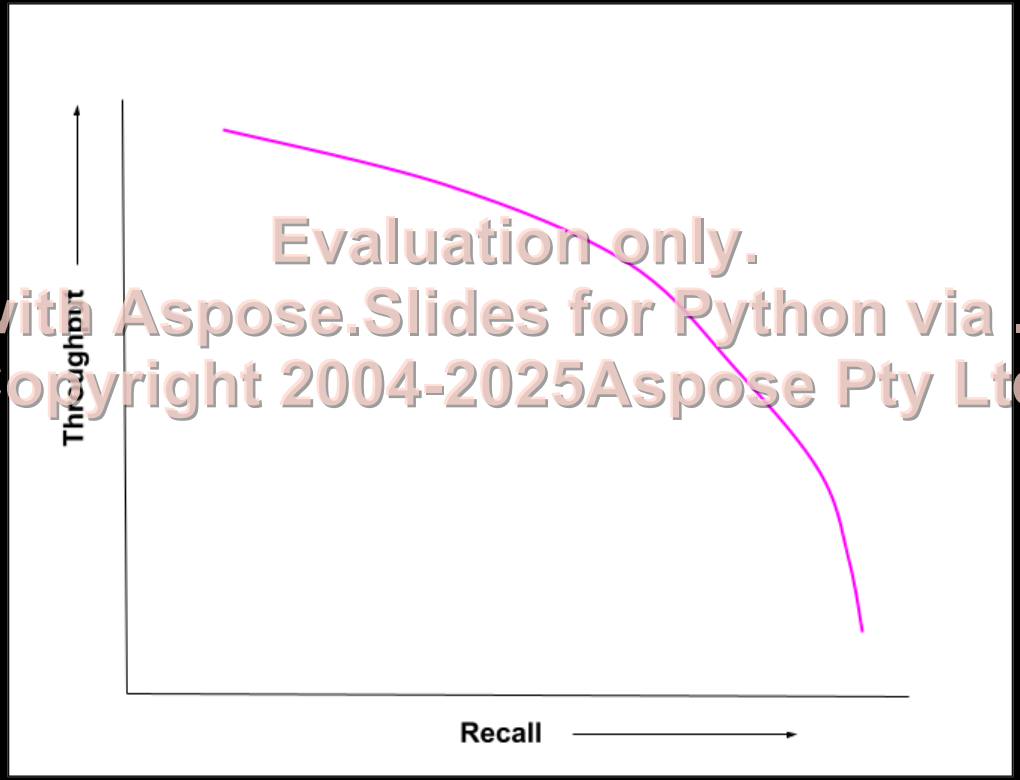
SMILE IDENTITY

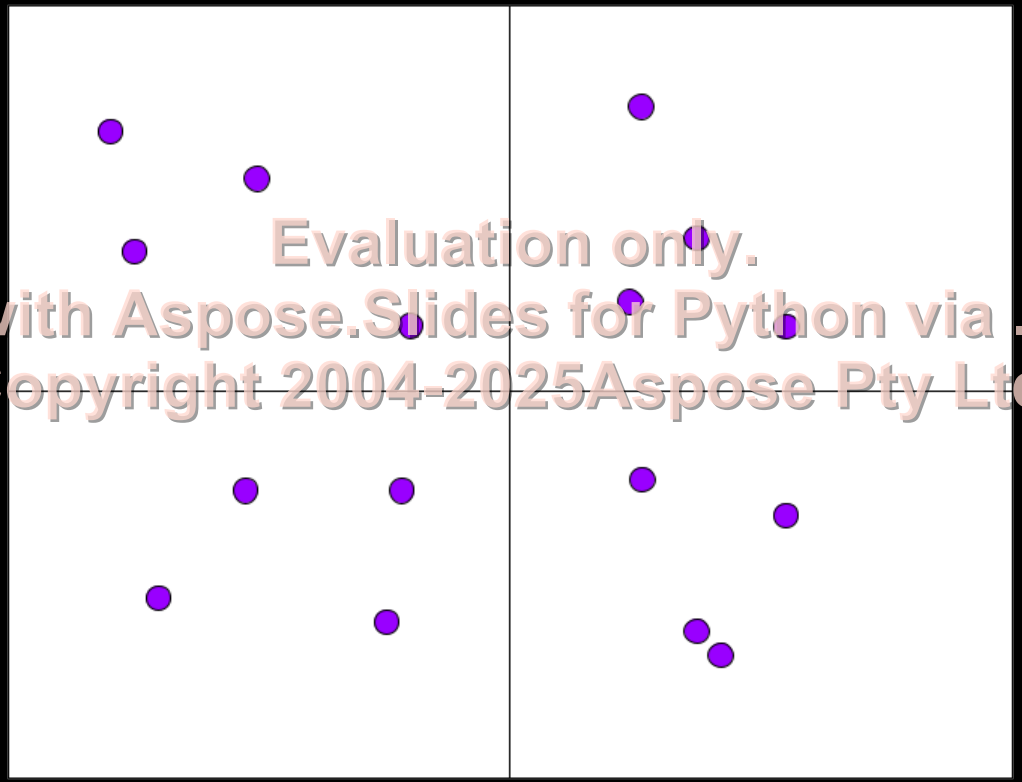
SMILE IDENTITY

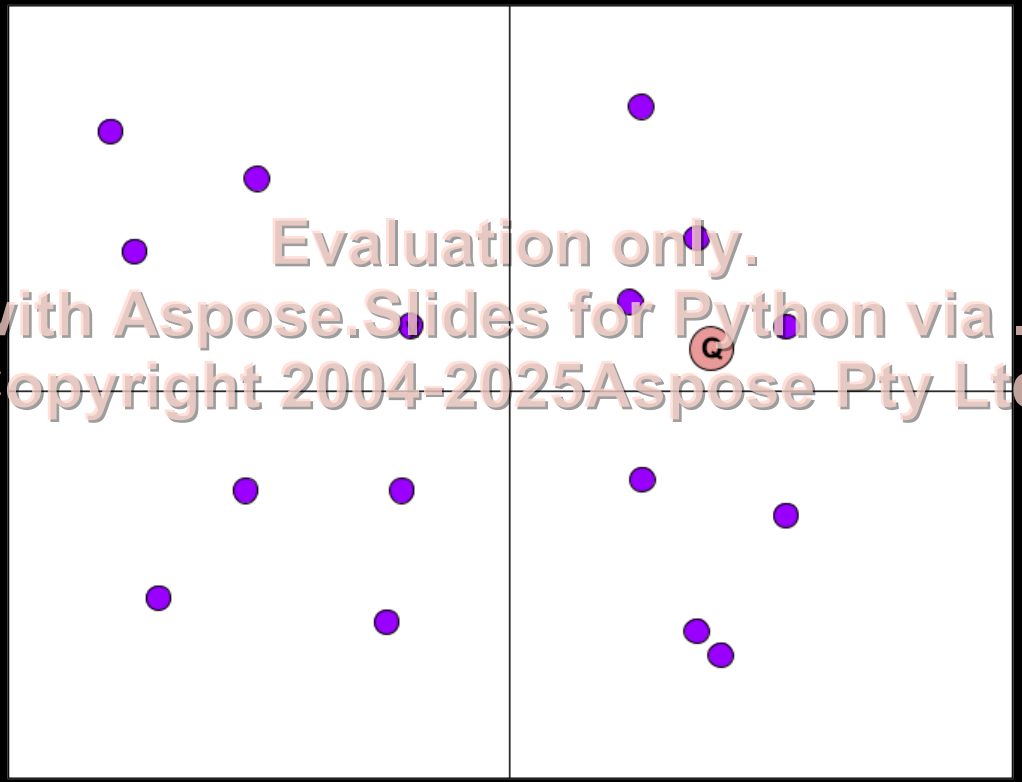
“This Track Is Wild!”

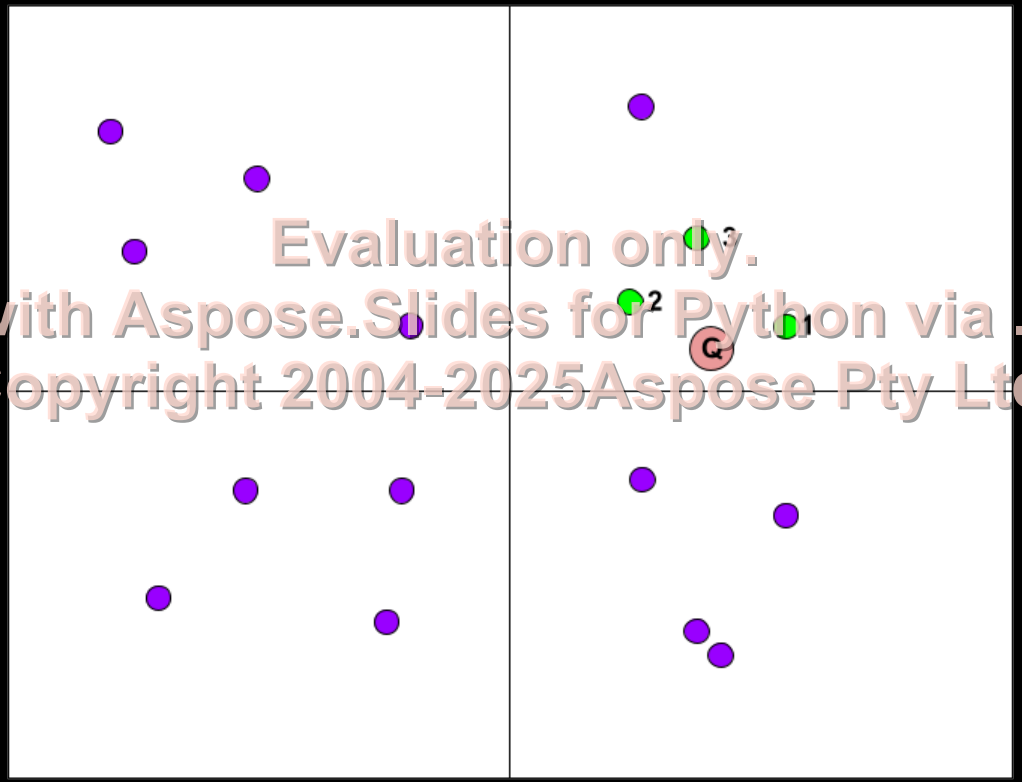
SMILE IDENTITY

SMILE IDENTITY

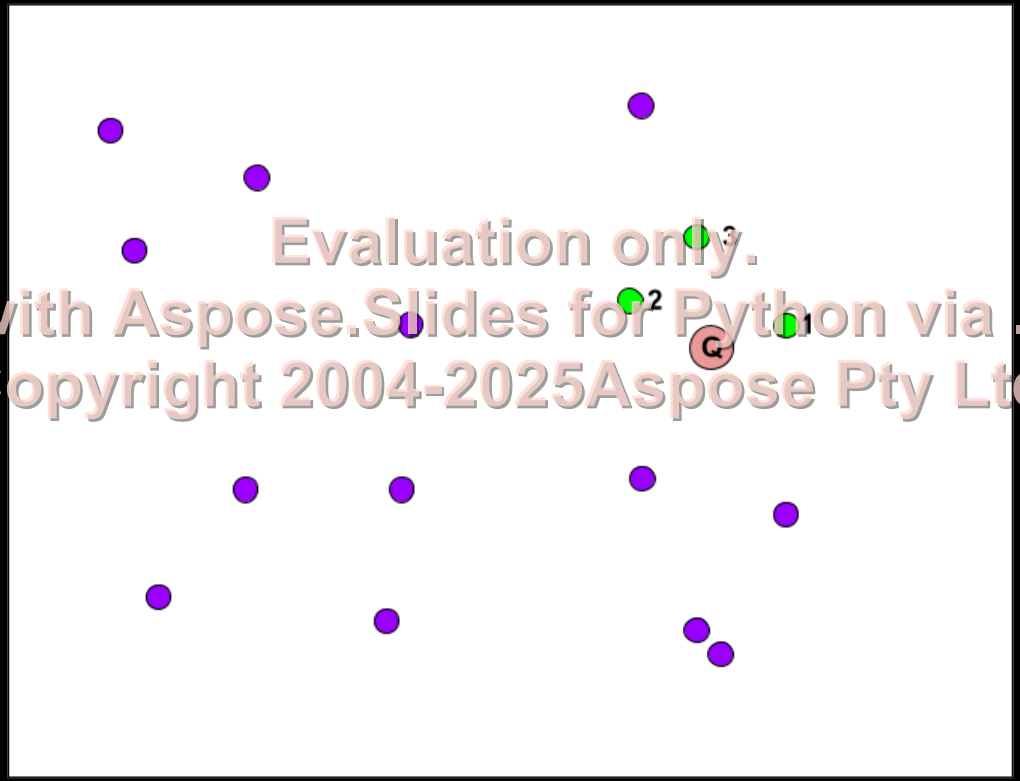






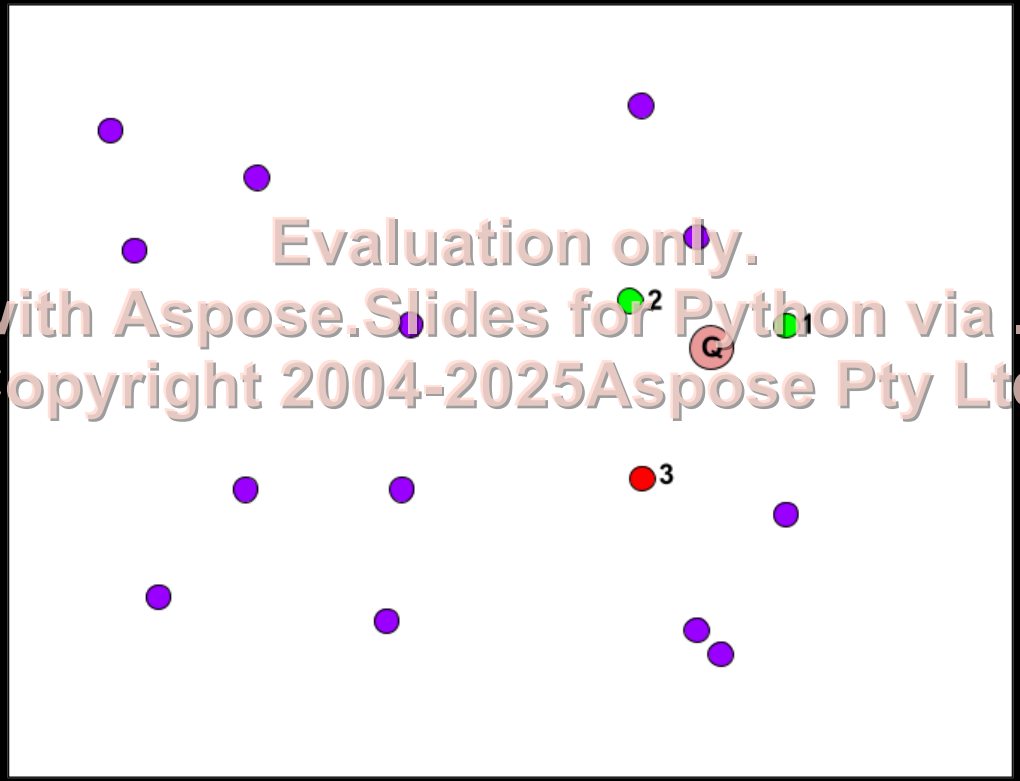


N = 3, Recall = 3/3



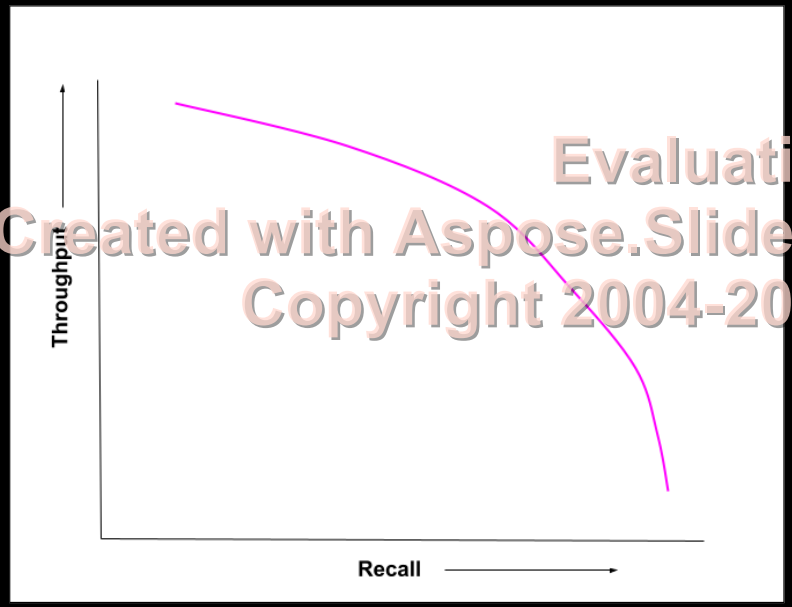
N = 3, Recall = 3/3

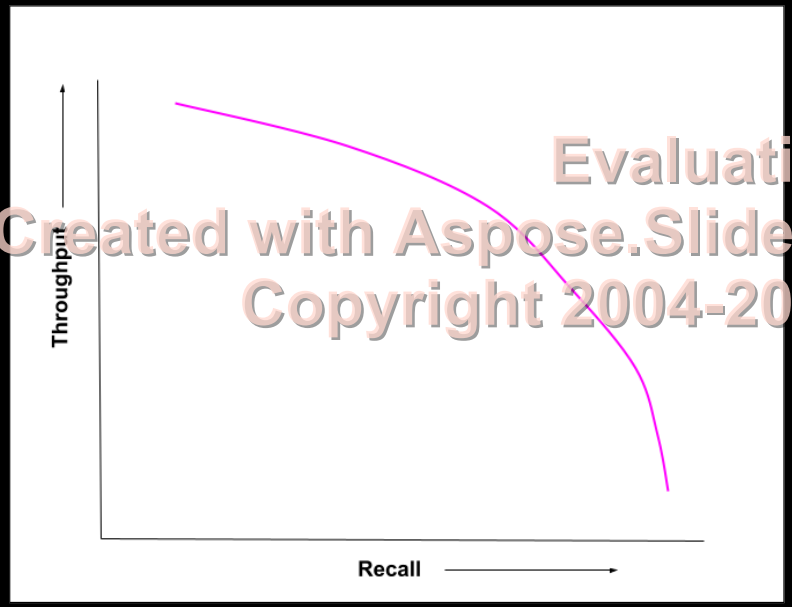
N = 3, Recall = 3/3

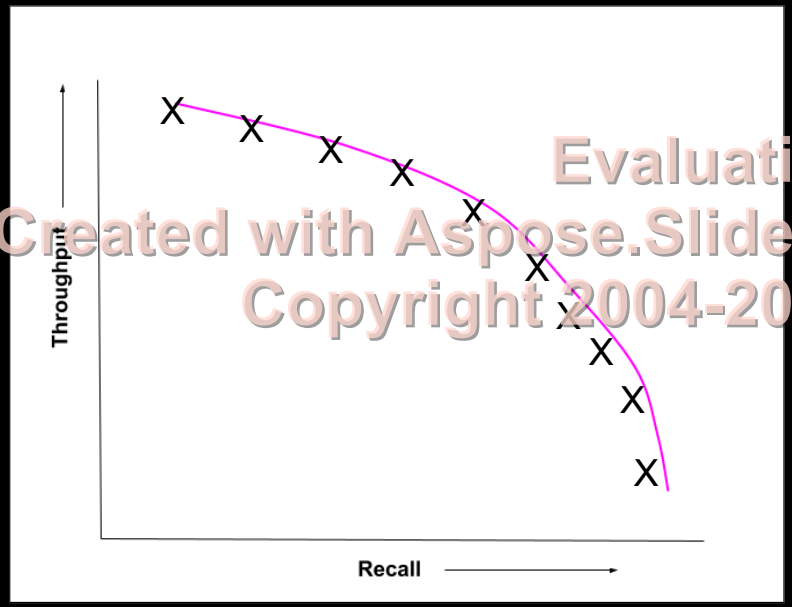


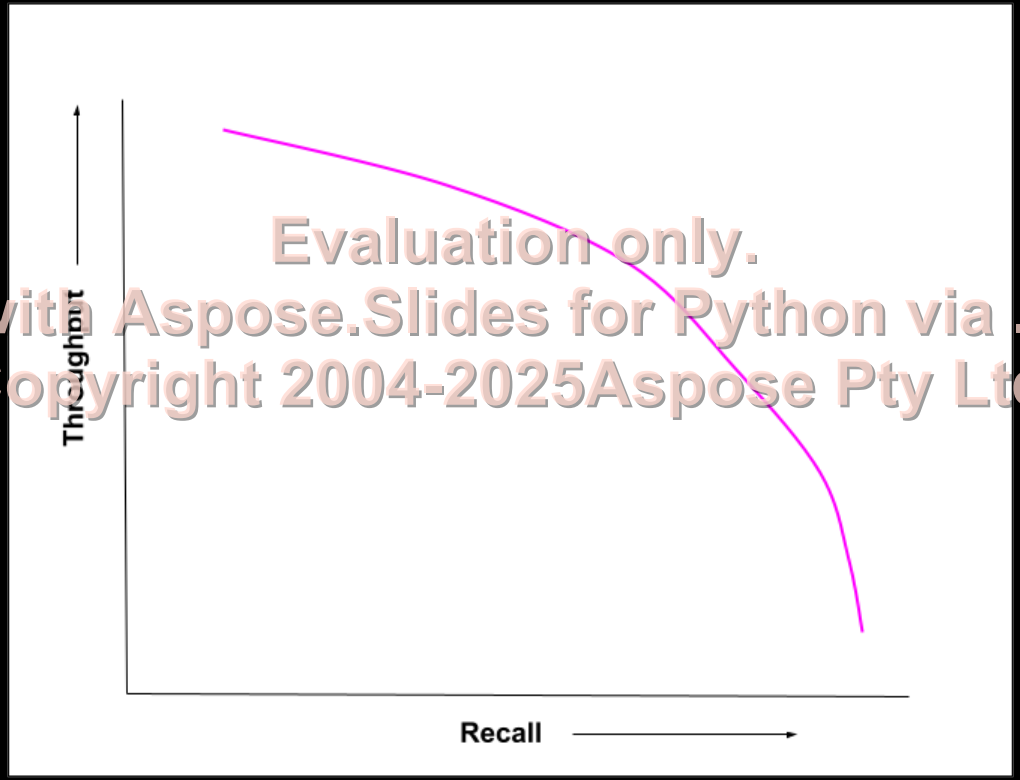
N = 3, Recall = 2/3

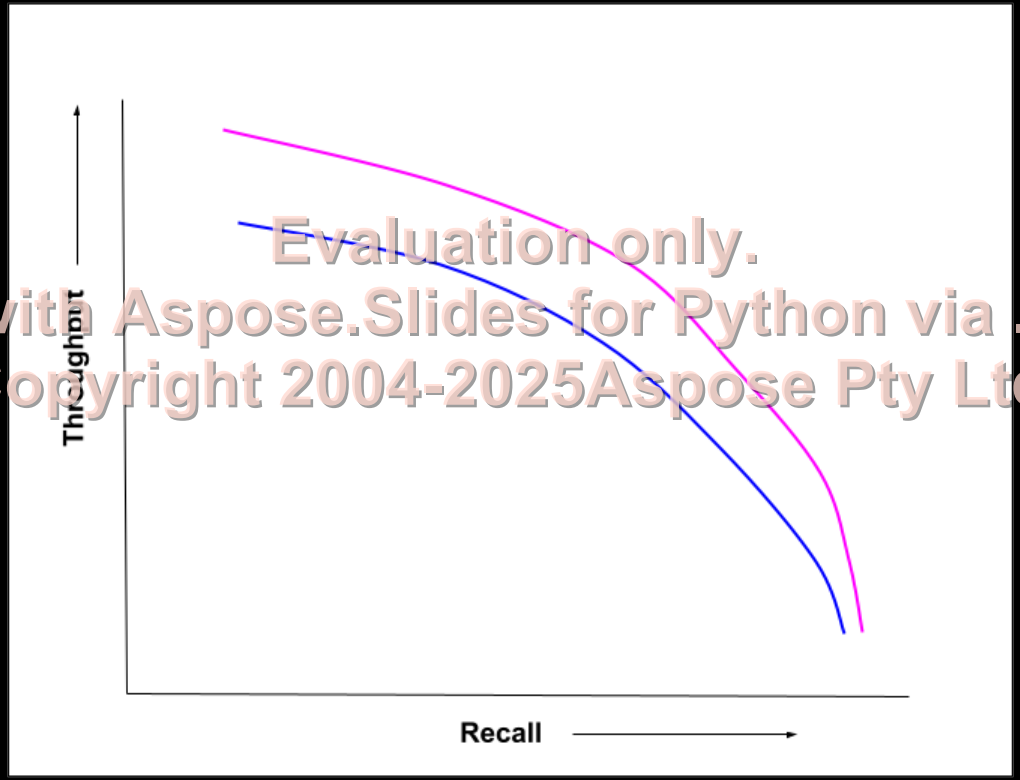
N = 3, Recall = 2/3

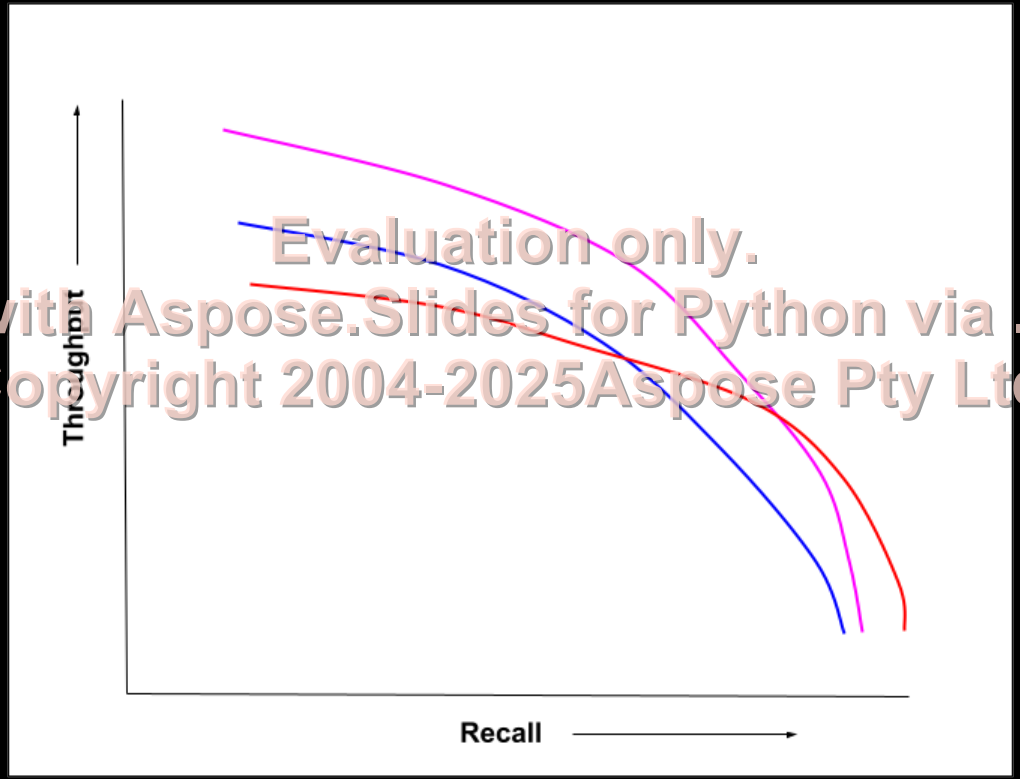


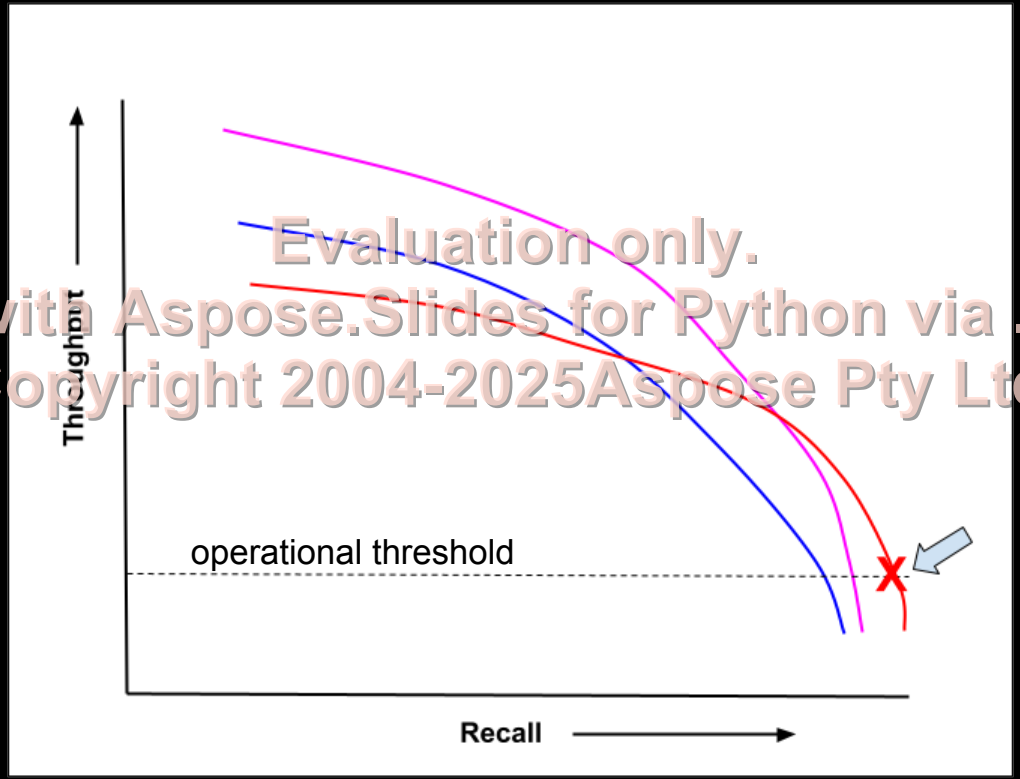












SMILE IDENTITY

Track	Operational Threshold
#1	10,000 QPS
#2	1,500 QPS
#3	2,000 QPS

SMILE IDENTITY

Track	Baseline Algorithm
#1	Faiss-IVFPQ
#2	MS-DiskANN
#3	Faiss-IVFPQ + 1 GPU

* Baseline algorithms provided by organizers
* Final ranking based on cumulative improvements over baseline

SMILE IDENTITY

Track	Benchmark	Winners
#1	Recall@10	Kuaishou+Tsinghua
#2	Recall@10	Zilliz+SUST
#3	Recall@10	Intel on Optane

SMILE IDENTITY

Winner: Kuaishou+Tsinghua U (15 participants)

Team	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
K+T U	0.711	0.723	0.7645	0.756	-	-

* threshold = 10,000 QPS

Winner: Kuaishou+Tsinghua U (15 participants)

Team	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
K+T U	0.711	0.723	0.7645	0.756	-	-
Faiss-IVFPQ	0.635	0.650	0.729	0.704	*0.069*	0.754

* threshold = 10,000 QPS

Winner: Zilliz+SUST (5 Participants)

Algo	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
Zilliz+SUST	–	0.760	–	0.495	–	*0.886*

* threshold = 1,500 QPS

Winner: Zilliz+SUST (5 Participants)

Algo	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
Zilliz+SUST	–	–	0.760	–	0.495	*0.886*
MS-DISKANN	0.950	0.937	0.900	0.936	0.489	*0.163*

* threshold = 1,500 QPS

Winner: Intel Optane (8 Participants)

Algo	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
Intel on Optane	0.9998	0.9998	0.9984	0.9957	0.9973	-

* threshold = 2,000 QPS

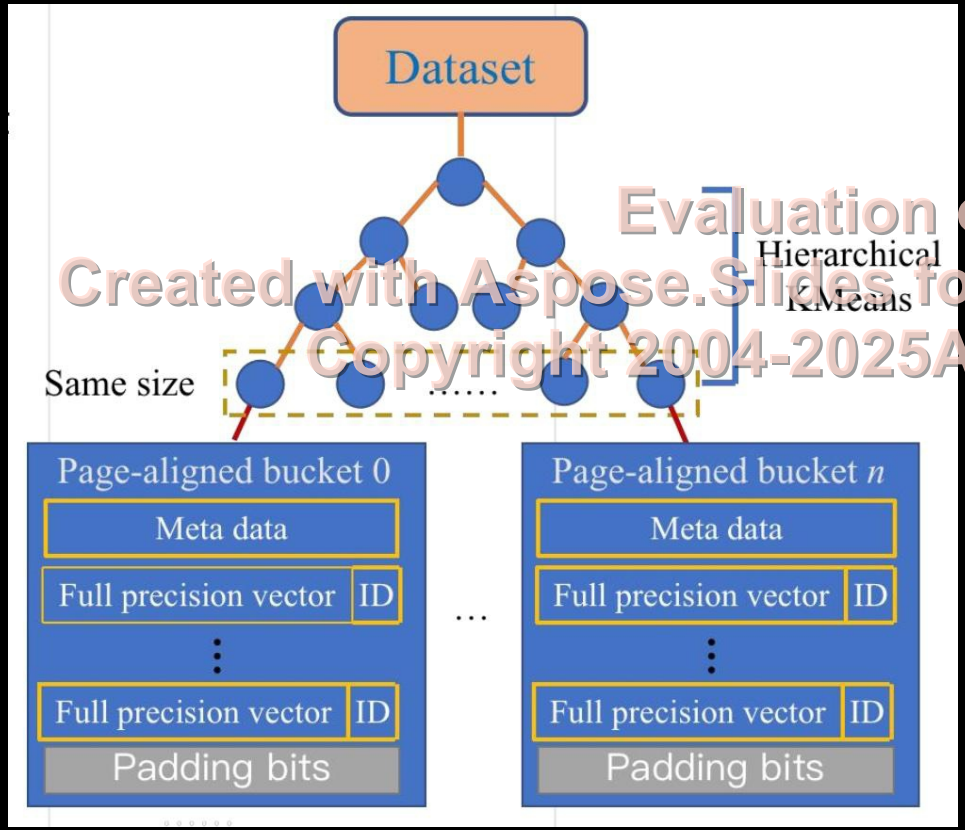
Winner: Intel Optane (8 Participants)

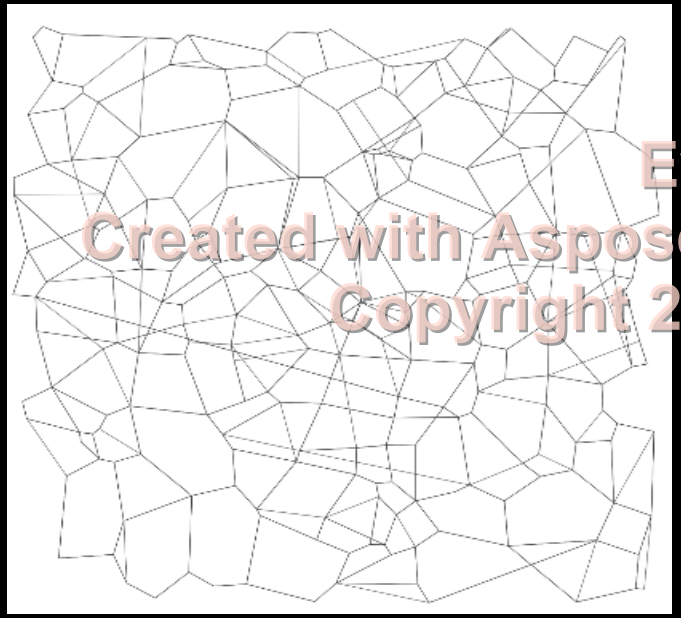
Algo	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
Intel on Optane	0.9998	0.9998	0.9984	0.9957	0.9734	-
Faiss-IVFPQ on GPU	0.9326	0.9428	0.9085	0.9132	0.8603	0.9786

* threshold = 2,000 QPS

Track	Benchmark	Winners
#1	Recall@10	Kuaishou+Tsinghua U
#2	Recall@10	ZilliZ+SJTU
#3	Recall@10	Intel on Optane
#3	Throughput (qps)	NVidia on 8 GPUs
#3	Power (KwH / q)	Intel on Optane
#3	Cost ($ / 100K qps * 4yr)	Intel on Optane

SMILE IDENTITY





DiskANN Design

DRAM

Compressed vectors
(50-60GB/billion vectors)

SSD

Low-diameter graph
+full-precision vectors

graph construction similar to HNSW

SMILE IDENTITY

	Team	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
1	Intel	0.9998	0.9998	0.9984	0.9957	0.9973	-
2	Nvidia	0.9988	0.9954	0.9899	0.99443	0.9469	-

SMILE IDENTITY

	Team	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
1	Nvidia	747K	802K	840K	584K	-	-
2	Intel	336K	197K	158K	161K	17K	-

SMILE IDENTITY

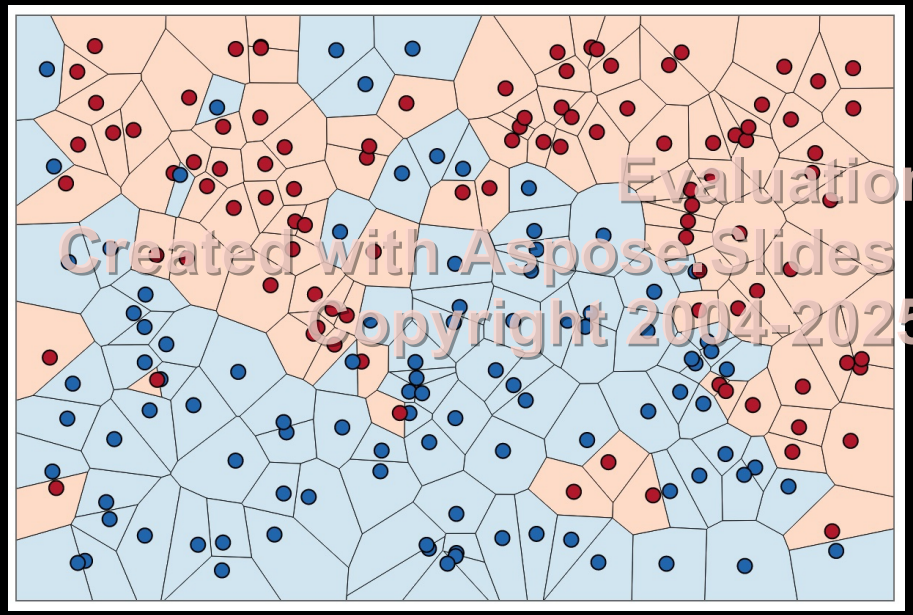
	Team	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
1	Intel	0.0022	0.0041	0.0049	0.0048	0.0446	-
2	Nvidia	0.0119	0.0112	0.0090	0.0090	0.0480	-

* (watt-sec) / q shown here

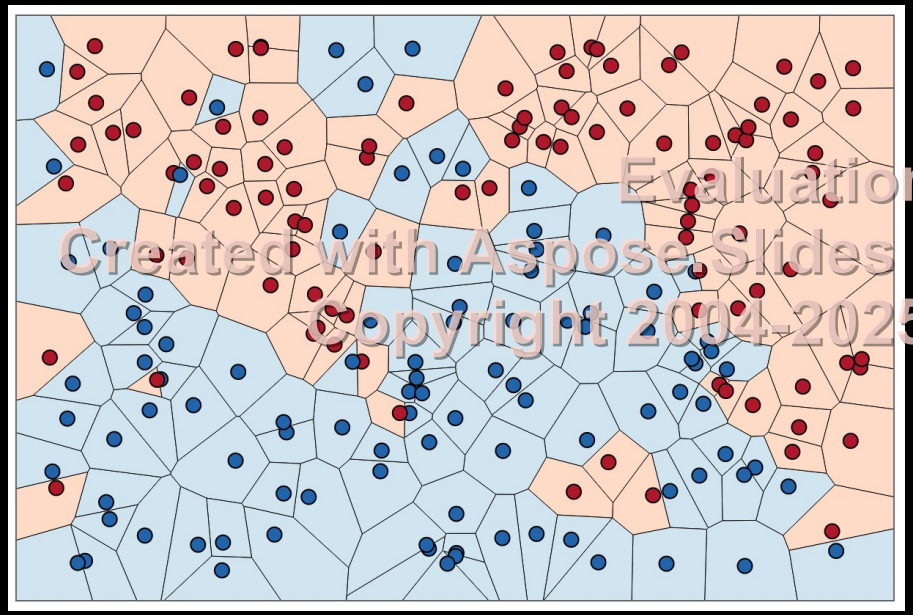
SMILE IDENTITY

	Team	BigANN	Deep	MSSpaceV	MSTuring	Text2Image	SSNPP
1	Intel	$15K	$16K	$16K	$16K	$104K	-
-	Nvidia	$304K	$304K	$153K	$153K	$917K	-

SMILE IDENTITY

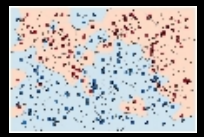


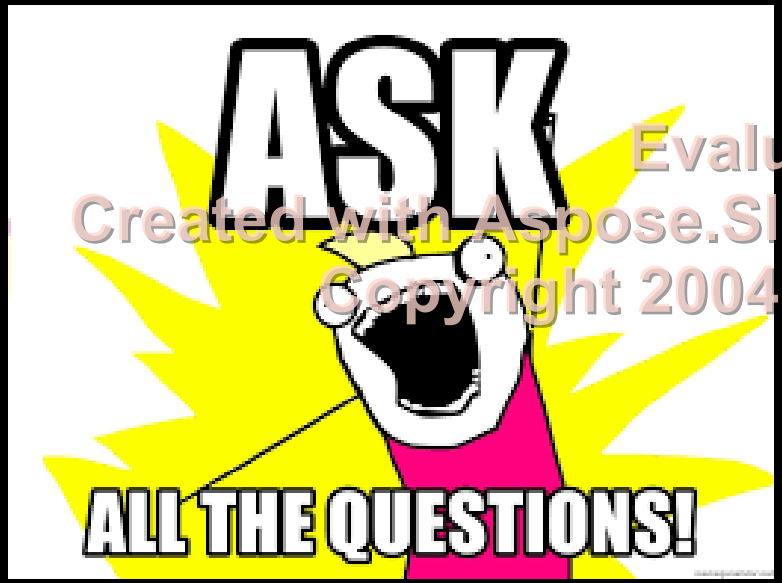
SMILE IDENTITY



SMILE IDENTITY

SMILE IDENTITY





SMILE IDENTITY