CSC 84040: Advanced Data Mining *with AI Agents*

Lecture 06: *Clustering, Automated Hypothesis Generation Based on Mining of the Literature, and Vision Models (SAM)*

By Saptarashmi Bandyopadhyay
Email: sbandyopadhyay@ccny.cuny.edu
Assistant Professor of Computer Science
City College of New York and Graduate Center at the City University of New York

October 7, 2026 CSC 84040

Saptarashmi Bandyopadhyay

Introduction

Massive scale breaks standard metric assumptions: high-dimensional geometry causes distances to concentrate, while petabyte-scale corpora render in-memory point assignments impossible.

Clustering serves as a foundation for discovery by allowing autonomous systems to **generate testable scientific hypotheses** and organize sparse biomedical knowledge graphs.

Modern computer vision reframes representation learning and spatial scene parsing as **online optimal transport and interactive proximity clustering**.

Saptarashmi Bandyopadhyay

Lecture Agenda

Massive-Scale & Streaming Clustering

01

Why high dimensions break geometric intuition, how BFR and CURE bypass RAM limits using summary statistics, and how BDMO tracks clusters across sliding streams.

Hypothesis Clustering & Literature-Based Discovery

02

Mining unstructured publications into dense semantic vectors, arranging them into typicality trees, and diffusing labels across graph Laplacians to discover novel science.

Online Clustering of Visual Representations

03

Replacing costly pairwise contrastive feature comparisons with online swapped prototype predictions regularized by entropy-constrained optimal transport.

Proximity Agents & Large Vision Models

04

Analyzing the Segment Anything Model (SAM), treating open-world image segmentation as prompt-conditioned spatial clustering backed by multi-hypothesis prediction heads.

Saptarshi Bandyopadhyay

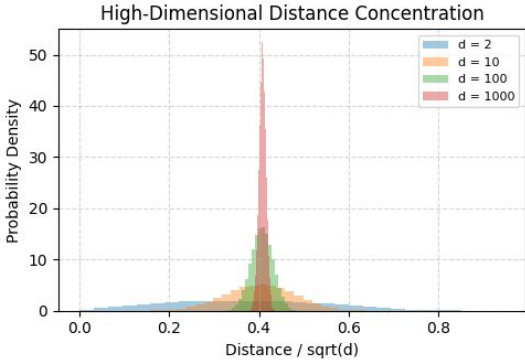
Massive-Scale & Streaming Clustering

Saptarashmi Bandyopadhyay

The Curse of Dimensionality (Clustering)

As dimension *d* grows large, pairwise Euclidean distances between random points **concentrate tightly around their expected value**, causing the relative contrast between near and far neighbors to vanish:

$$\lim_{d \to \infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} = 0$$



Here, $\text{dist}_{\max}$ and $\text{dist}_{\min}$ denote the maximum and minimum pairwise Euclidean distances observed among points uniformly distributed in the *d*-dimensional unit hypercube.

The dot product between two independent zero-mean vectors has expectation 0, meaning that **random vectors in high dimensions are approximately orthogonal** with probability approaching 1.

When displacement vector (*p* - *c*) is roughly perpendicular to (*c* - *q*), the distance between distant points can be approximated via the Pythagorean relation $d(p, q)^2 \approx d(p, c)^2 + d(c, q)^2$

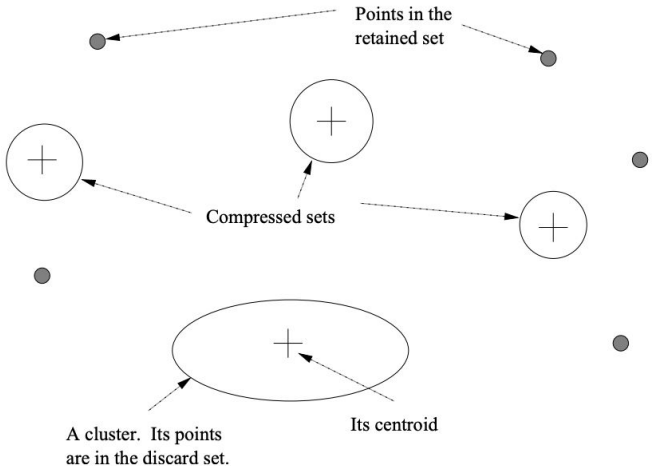
Santanshini Bandyopadhyay

Memory-Constrained Clustering with Bradley-Fayyad-Reina

Iterating over terabytes of training data using standard *k*-means causes **severe disk thrashing** when dataset scale exceeds physical RAM.

The Bradley-Fayyad-Reina (BFR) algorithm scales clustering by assuming clusters are **normally distributed along axis-aligned ellipsoids**, allowing dimensions to be evaluated independently.

Under BFR, data points stream through memory in single-pass chunks; once assigned to a cluster, the raw coordinate vectors are **immediately purged from main memory**.



Saptarshi Bandyopadhyay

BFR Representation: Sufficient Statistics

Instead of storing point coordinates, BFR summarizes each cluster using **2*d* + 1 running accumulators**: count *N*, coordinate sum vector *SUM*, and squared sum vector *SUMSQ*:

$$N = \sum_{j=1}^{N} 1, \quad SUM_i = \sum_{j=1}^{N} x_{ji}, \quad SUMSQ_i = \sum_{j=1}^{N} x_{ji}^2$$

Here, *N* is the number of points in the cluster, x_ji is the coordinate of point *j* along dimension *i*, and *d* is the total dimensionality of the feature space.

These accumulators yield the coordinate centroid $c_i = SUM_i/N$ and dimension-wise coordinate variance $\sigma_i^2 = (SUMSQ_i/N) - (SUM_i/N)^2$ in *O*(1) time.

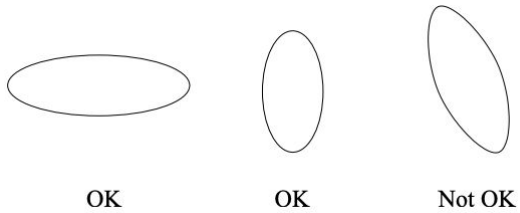
Merging clusters or adding new points reduces to **simple vector addition** ($N_{\text{new}} = N_1 + N_2$, $SUMSQ_{\text{new}} = SUMSQ_1 + SUMSQ_2$, $SUM_{\text{new}} = SUM_1 + SUM_2$), eliminating the need to re-read earlier points from disk.

Saptarashmi Bandyopadhyay

BFR Assignment: Mahalanobis Distance Gating

Standard Euclidean distance distorts cluster boundaries when dimensions have unequal variance; BFR corrects for this by scaling distance with cluster variance using the Mahalanobis metric:

$$D_M(x, c) = \sqrt{\sum_{i=1}^{d} \left(\frac{x_i - c_i}{\sigma_i} \right)^2}$$



Here, x_i is the coordinate of candidate point *x* along dimension *i*, c_i is the centroid coordinate, and σ_i is the standard deviation of the cluster along dimension *i*.

For normally distributed coordinates, D_M(*x*, *c*) follows a χ-distribution; points are assigned to a cluster only if their distance falls below a statistical gating threshold τ ≈ 3σ to 4σ.

Points failing this check are clustered together in memory: cohesive groups form Compressed Set miniclusters, while truly isolated points remain in the Retained Set.

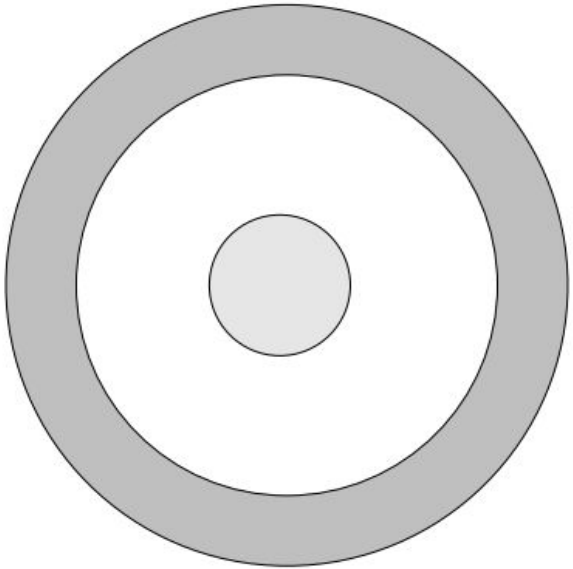
Saptarshi Bandyopadhyay

The CURE Algorithm

When clusters form concentric rings (see Figure), elongated ribbons, or interlocking spirals, single-centroid methods fail because **the geometric center lies in empty space**.

CURE (Clustering Using REpresentatives) resolves this by capturing the shape of a cluster using a **collection of *c* well-dispersed boundary points** rather than a single centroid. Here, *c* represents a small user-defined number of representative points tracking the geometric perimeter of the cluster.

This multi-point representation allows CURE to adapt to arbitrary, non-spherical geometries without relying on rigid parametric or Gaussian assumptions.



Saptarshi Bandyopadhyay

CURE Mechanics: Dispersing and Shrinking

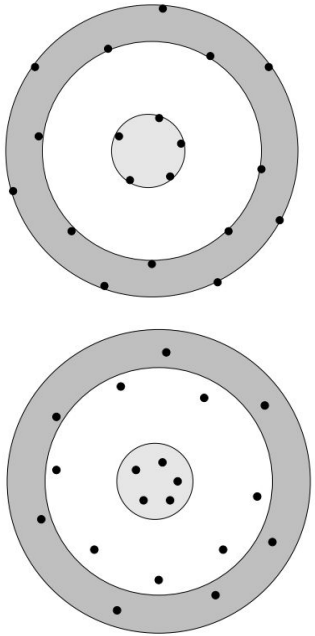
On a random sample of the dataset, CURE selects *c* representative points per cluster by choosing points that are **maximally distant from one another** (Figure 1) to trace the cluster's perimeter.

To prevent isolated noise points from bridging distinct clusters, representatives are **contracted toward the centroid** by a shrinkage fraction α ≈ 0.2 (Figure 2):

$$p' = p + \alpha(c_{\text{mean}} - p)$$

Here, *p* is the original representative coordinate, c_mean is the mean centroid of the sample cluster, and p' is the resulting contracted representative location.

After initial clustering in memory, the full dataset streams from disk: each point is assigned to the cluster of its **nearest contracted representative point**.



Sanjaya Mi Bandyopadhyay

Stream Clustering Over Sliding Windows: BDMO

Streaming applications require tracking clusters over a sliding window of the last N points, where historical records **cannot be stored permanently**.

Adapting the DGIM bit-counting algorithm, the BDMO framework organizes the stream into buckets whose sizes grow as **powers of two** (p, $2p$, $4p$, ...), holding at most two buckets of any given size.

Each bucket stores timestamps and summary statistics (N, centroid/clustroid, dispersion); when three buckets share a size, the oldest two **merge hierarchically**.

This bucketing scheme bounds total memory use to $O(k \log N)$, allowing queries over the last m points to be answered in sublinear time.

Sahtarashmi Bandyopadhyay

Use Case: Clustering for Mining of Scientific Hypotheses

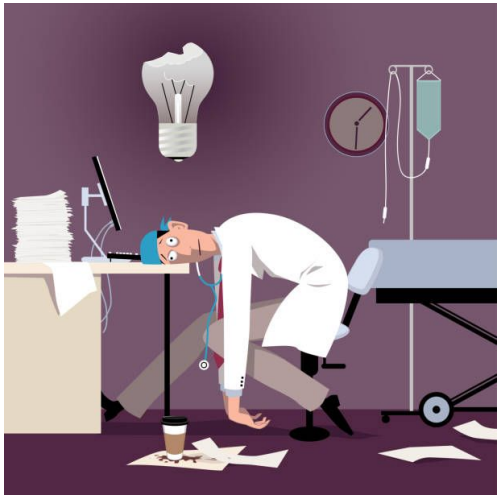
Saptarashmi Bandyopadhyay

Literature Mining for Scientific Hypotheses

Human reading capacity cannot keep pace with the accelerating volume of science, where over 50 million manuscripts exist and a new article is indexed every 30 seconds.

In molecular oncology, critical targets like the tumor suppressor protein p53 accumulate over 70,000 papers, creating an **insurmountable backlog** for bench scientists.

The Knowledge Integration Toolkit (KnIT) (Spangler et. al. 2014) overcomes this cognitive limit by **projecting unstructured scientific text into metric spaces**, turning document mining into an inductive engine for hypothesis generation.



Jyotiradhmi Bandyopadhyay

Building the Semantic Feature Space

We gather Medline abstracts referencing each human kinase while **excluding co-mentions of p53** to eliminate trivial lexical shortcuts and circular associations.

Document abstracts are vectorized across the top 20,000 words and bi-grams using Term Frequency-Inverse Document Frequency (TF-IDF) weighting combined with Euclidean unit-norm scaling:

$$\text{TF-IDF}_{i,j} = \frac{f_{i,j}}{\max_k f_{k,j}} \cdot \log_2 \left(\frac{|\mathcal{D}|}{n_i} \right)$$

Here, f_i,j is the frequency of word or phrase *i* in abstract *j*, the denominator max_k f_k,j is the maximum frequency of any term in abstract *j*, $|\mathcal{D}|$ is the total number of document abstracts, and n_i is the number of documents containing term *i*.

- Averaging all normalized vectors mentioning a specific kinase produces a **kinase centroid**, embedding its functional cellular context within an interpretable coordinate space.

Saptarashmi Bandyopadhyay

Relative Neighborhood Trees

Standard network visualizations often degenerate into dense, uninterpretable "hairballs" when every entity is linked based on continuous similarity.

KnIT resolves this by building an *n*-ary tree initialized at the **most typical entity**—the specific kinase whose vector lies closest to the global dataset centroid.

Unassigned entities attach greedily to the nearest tree node under branching constraints, ensuring that **increasing tree depth reflects specialized functional variation** rather than general background activity. Without receiving any biochemical supervision, known p53 kinases **cluster spontaneously into localized subtrees**, highlighting neighboring uncharacterized nodes as candidate regulators.

Sabarashmi Bandyopadhyay

Prioritizing Hypotheses via Graph Laplacian Diffusion

- Given a sparse binary vector $y \in \{0, 1\}^n$ marking whether kinase *i* is known to phosphorylate p53, we seek continuous affinity scores $f \in \mathbb{R}^n$ over the 10-nearest-neighbor network.
- The semi-supervised diffusion objective balances **fidelity to validated historical facts** against **smoothness across graph edges** using the unnormalized Laplacian matrix $L = D - A$:

$$\min_f \left\{ (f - y)^T (f - y) + \mu f^T L f \right\}$$

$$f = (I + \mu L)^{-1} y$$

Here, *y* is the initial binary label vector, *f* is the diffused continuous ranking vector, *I* is the identity matrix, *A* is the symmetric adjacency matrix, *D* is the diagonal degree matrix with $D_{ii} = \sum_j A_{ij}$, *L* is the graph Laplacian ($L = D - A$), and μ is the diffusion regularization coefficient.

- Setting the parameter μ to the inverse of the maximum row sum of *L* guarantees a strictly positive definite Hessian, ensuring **convexity and a unique global optimum**.

Sanjarashmi Bandyopadhyay

Predicting Future Discoveries

To test whether automated text clustering predicts true biological discoveries, the input literature was frozen to **abstracts published before 2003**, when only 10 p53 kinases had been discovered.

Label diffusion across the historical literature network successfully **identified 7 of the 9 kinases discovered after 2003** near the top of the candidate ranking, yielding an AUC of 0.840.

In leave-one-out tests across 401 kinases, literature-derived metric vectors achieved an AUC of 0.800, **outperforming standard biological sequence alignment** algorithms (BLAST, AUC = 0.751

Saptarashmi Bandyopadhyay

Laboratory Verification of Novel Kinases

Rather than relying solely on retrospective cross-validation, KnIT's top uncharacterized kinase predictions were subjected to **direct experimental biochemical assays**.

In vitro electrophoresis assays tracking radioactive phosphate transfer (γ-32P-ATP) confirmed that **top candidates NEK2 and PKN1 directly phosphorylate p53**, while low-ranked controls showed no activity.

Cellular co-immunoprecipitation assays verified that NEK2 and PKN1 **physically bind to p53 in living human cells**, establishing a verified link between text-based clustering and genuine laboratory discovery.

Saptarashmi Bandyopadhyay

Online Clustering of Visual Representations

Saptarashmi Bandyopadhyay

The Contrastive Pairwise Bottleneck in Visual Learning

Standard contrastive self-supervised algorithms treat every image instance as its own discrete class, attempting to pull augmented views together while pushing all other images away.

Computing these pairwise negative distances creates a steep scaling hurdle, requiring **thousands of negative feature vectors per step** via massive batch sizes (4096+) or large memory banks (65k+ vectors).

SwAV (Swapping Assignments between Views) circumvents pairwise feature comparison by **mapping representations to a set of trainable prototypes**, converting instance discrimination into online cluster assignment.

Saptarashmi Bandyopadhyay

The SwAV Objective: Swapped Prediction Across Views

- Given augmented views x_t and x_s of the same image, visual encoder f_θ extracts normalized feature embeddings z_t, z_s ∈ $\mathbb{R}^D$ and maps them to cluster prototype codes q_t, q_s.
- Invariance is enforced through a swapped prediction problem: the network must **predict the cluster assignment of one view using the visual feature of the other view**:

$$\mathcal{L}(z_t, z_s) = \ell(z_t, q_s) + \ell(z_s, q_t)$$

$$\ell(z_t, q_s) = -\sum_{k=1}^{K} q_s^{(k)} \log p_t^{(k)}, \quad p_t^{(k)} = \frac{\exp\left(\frac{1}{\tau} z_t^T c_k\right)}{\sum_{k'=1}^{K} \exp\left(\frac{1}{\tau} z_t^T c_{k'}\right)}$$

Here, z_t, z_s are unit-norm feature vectors, q_t, q_s ∈ [0, 1]^K represent cluster assignment codes, c_k ∈ $\mathbb{R}^D$ is the *k*-th trainable prototype vector, τ is a temperature parameter, and p_t^(k) is the predicted softmax probability of assigning feature z_t to prototype c_k.

- Prototype vectors are updated online via backpropagation alongside the backbone encoder, **eliminating the need for separate offline clustering phases**.

Soutarashmi Bandyopadhyay

The Equipartition Constraint

- Unconstrained online clustering easily collapses into trivial degenerate representations where the visual encoder projects every image onto a **single dominant prototype vector**.SwAV prevents representational collapse by imposing an **equipartition constraint** that forces the mini-batch examples to partition equally across all *K* prototypes.
- Finding the optimal code assignment is formulated as an entropy-regularized optimal transport problem over the transportation polytope $\mathcal{Q}$:

$$\max_{Q \in \mathcal{Q}} \operatorname{Tr}(Q^T C^T Z) + \epsilon H(Q)$$

$$\mathcal{Q} = \left\{ Q \in \mathbb{R}_+^{K \times B} \;\middle|\; Q\mathbf{1}_B = \frac{1}{K}\mathbf{1}_K, \quad Q^T\mathbf{1}_K = \frac{1}{B}\mathbf{1}_B \right\}$$

Here, $Z \in \mathbb{R}^{D \times B}$ denotes the batch feature matrix of B representations in dimension D, $C \in \mathbb{R}^{D \times K}$ is the matrix of K trainable prototypes, $Q \in \mathbb{R}_+^{K \times B}$ is the continuous assignment matrix, $H(Q) = -\sum_{ij} Q_{ij} \log Q_{ij}$ represents the entropy function, ε controls mapping smoothness, and 1_K, 1_B are vectors of ones.

- Requiring each prototype to be selected on average at least *B* / *K* times per batch prevents prototype starvation without needing offline reassignment steps.

Sbtarashmi Bandyopadhyay

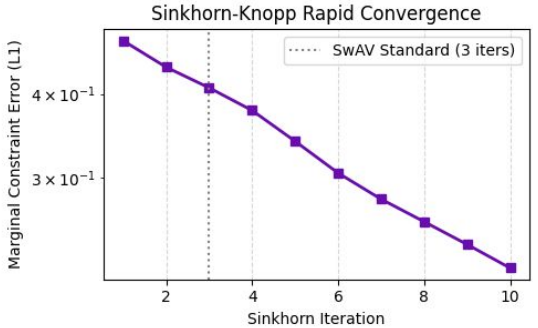
Online Soft Clustering via Sinkhorn-Knopp

- Rounding cluster assignments to discrete integer codes causes aggressive gradient jumps that destabilize representation learning during mini-batch SGD.
- SwAV preserves continuous soft assignment codes Q*, computing the regularized optimal transport solution through **iterative dual vector scaling**:

$$Q^* = \mathrm{Diag}(u)\exp\left(\frac{C^T Z}{\epsilon}\right)\mathrm{Diag}(v)$$

$$u \leftarrow \frac{\mathbf{1}_K/K}{\exp(C^T Z/\epsilon)v}, \quad v \leftarrow \frac{\mathbf{1}_B/B}{\exp\left(C^T Z/\epsilon\right)^T u}$$

Here, Q^* is the optimal soft assignment matrix, $u \in \mathbb{R}^K$ and $v \in \mathbb{R}^B$ are scaling vectors updated alternately, Diag(·) converts a vector into a diagonal matrix, and C, Z, ε retain their earlier definitions.

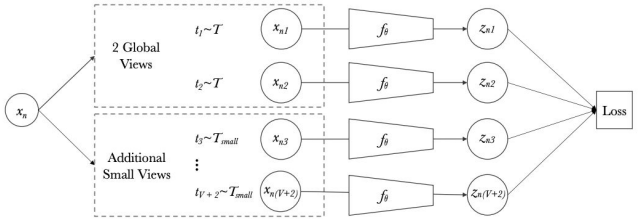


- In practice, **just 3 Sinkhorn iterations** achieve numerical convergence, adding only 35ms of GPU overhead for a mini-batch of 4,096 images.

Sattarasmi Bandyopadhyay

Multi-Cropping Augmentation

Comparing multiple crops per image improves learned visual features, but repeating full-resolution views (224 × 224) causes GPU memory and computation to scale quadratically.



- The multi-crop strategy resolves this by pairing **two high-resolution global views** (224 × 224) with *V* low-resolution local crops (96 × 96) that capture focused sub-regions.
- Prototype codes are generated exclusively from the global views to preserve holistic scene semantics, while local crops are tasked with predicting those global codes:

$$\mathcal{L}_{\text{multi-crop}} = \sum_{i \in \{1,2\}} \sum_{v=1}^{V+2} \mathbf{1}_{v \neq i} \ell(z_{t_v}, q_{t_i})$$

Here, index *i* ∈ {1, 2} denotes the two full-resolution global views, *v* indexes all *V* + 2 augmented views, z_{t_v} is the normalized representation of view *v*, q_{t_i} is the cluster assignment code from global view *i*, and 1_{v ≠ i} prevents trivial self-comparisons.

Saptarshi Bandyopadhyay

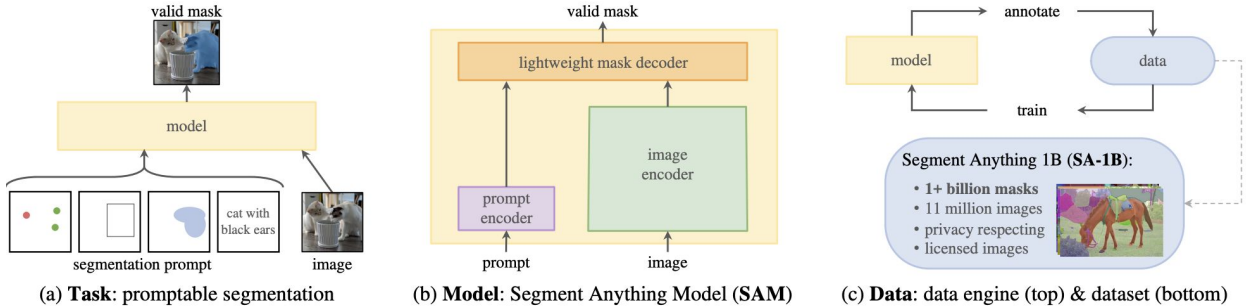
Proximity Agents & Vision Models

Saptarashmi Bandyopadhyay

Segmentation as Prompt-Driven Spatial Proximity Clustering

Computer vision has historically trained separate models for narrow, siloed tasks—such as edge detection, instance segmentation, and foreground extraction—each with rigid output semantics.

Framing segmentation as a general promptable task enables zero-shot transfer via prompt engineering: an external agent can perform edge discovery or instance tracking simply by supplying regular point grids or detector boxes.



Kirillov et al. (2023), *Segment Anything*, Figure 1

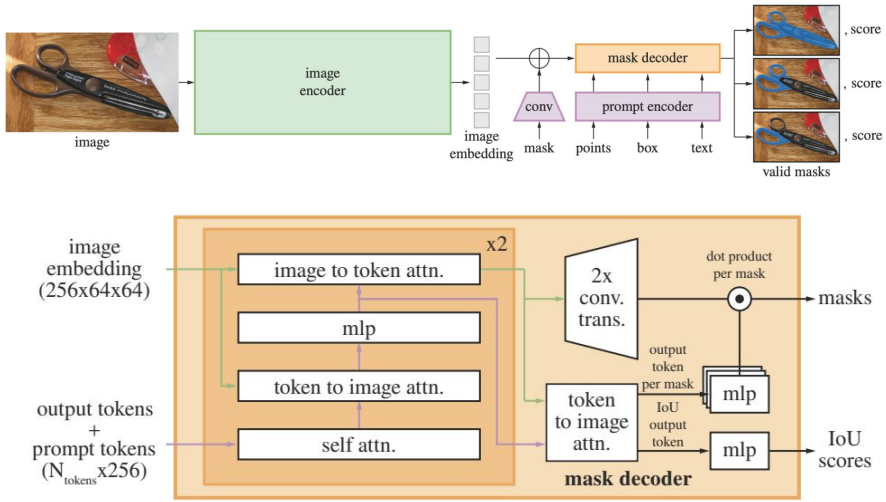
Rashmi Bandyopadhyay

Segment Anything Model (SAM) Architecture

A Vision Transformer (ViT-H/16) pre-trained with Masked Autoencoders processes the input image once into a 64 × 64 spatial embedding, **amortizing image encoding costs across hundreds of subsequent prompts**.

A lightweight two-layer Transformer decoder performs bidirectional cross-attention between prompt tokens and image embeddings, **predicting segmentation masks in ~50ms on CPU**.

Decoupling heavy visual feature extraction from fast prompt decoding allows interactive proximity agents to query scenes in real time directly within a web browser.



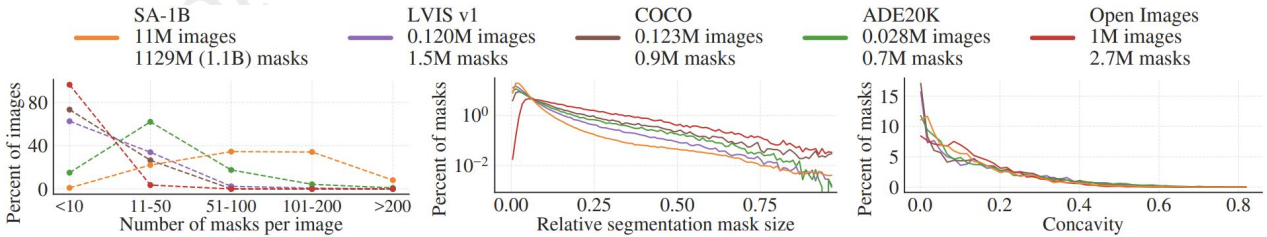
Kirillov et al. (2023), *Segment Anything*, Figure 4 & Figure 14.

The SA-1B Dataset

Model-in-the-Loop Data Collection: Annotation scaled across three iterative stages—assisted-manual labeling (4.3M masks), semi-automatic object proposal (5.9M masks), and fully automatic grid prompting (1.1B masks).

Dataset Scale (SA-1B): Contains **11 million licensed images and 1.1 billion masks**, providing 400× more masks than any prior segmentation dataset.

Empirical Mask Quality: Human validation confirms that 94% of automatically generated masks achieve over 90% IoU when compared against manual pixel corrections by professional annotators.



Kirillov et al. (2023), *Segment Anything*, Figure 6.

Sudeshna Bandyopadhyay

Summary of Today's Class

Clustering provides a unified mathematical foundation across diverse modalities:

- Summarizing massive streaming data into sublinear sufficient statistics (BFR/BDMO).
- Discovering hidden functional connections in sparse scientific literature spaces (KnIT).
- Mapping continuous visual representations to semantic prototypes via optimal transport (SwAV).
- Grouping spatial pixels into discrete object boundaries via interactive prompts (SAM).

Embodied agents in physical environments cannot afford exhaustive scene parsing; they rely on **prompt-driven proximity grouping** to attend selectively to task-relevant objects.
Open research challenges include **distilling heavy Vision Transformers into real-time edge devices** and extending visual grouping from appearance similarity to causal, physics-based affordances

Sankarshan Bandyopadhyay

Questions?

Saptarashmi Bandyopadhyay