Unsupervised Learning

Unsupervised Learning

- Supervised learning used labeled data pairs $(\mathbf{x}, \mathrm{y})$ to learn a function $f : \mathrm{X} \rightarrow \mathrm{Y}$
  - But, what if we don't have labels?

- No labels = **unsupervised learning**
- Only some points are labeled = **semi--supervised learning**
  - Labels may be expensive to obtain, so we only get a few

- **Clustering** is the unsupervised grouping of data points. It can be used for **knowledge discovery**.

Tasks

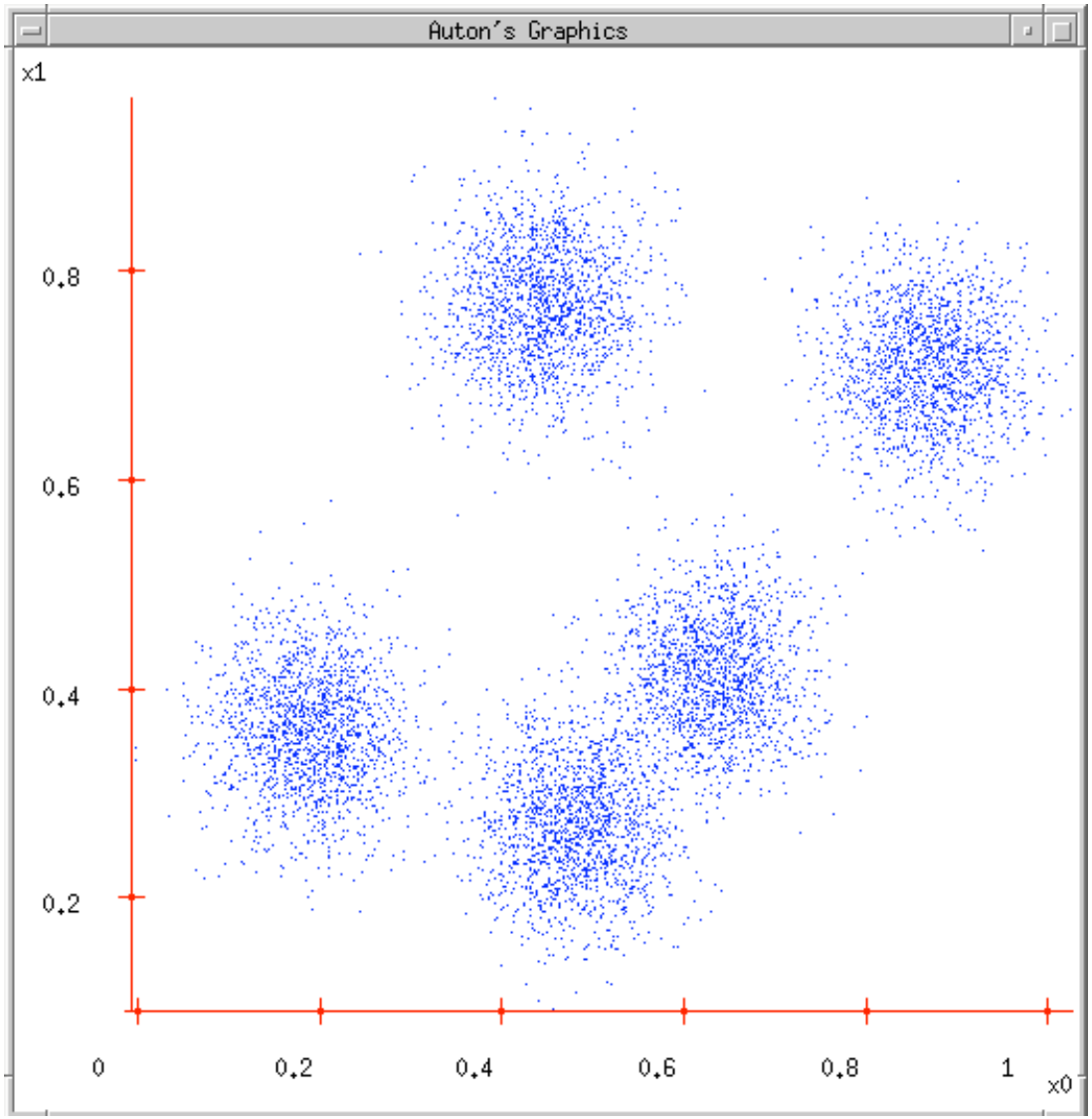
- Clustering
  - K-Means (*Covered today*)
- Dimensionality Reduction
  - PCA (*Not covered*)
- Density Estimation
  - Gaussian Mixture Models (*Not covered*)
- Generative Modeling
  - VAE, GAN (*Next lectures*)

K–Means Clustering

Some material adapted from slides by Andrew Moore, CMU.

Visit http://www.autonlab.org/tutorials/ for
Andrew's repository of Data Mining tutorials.

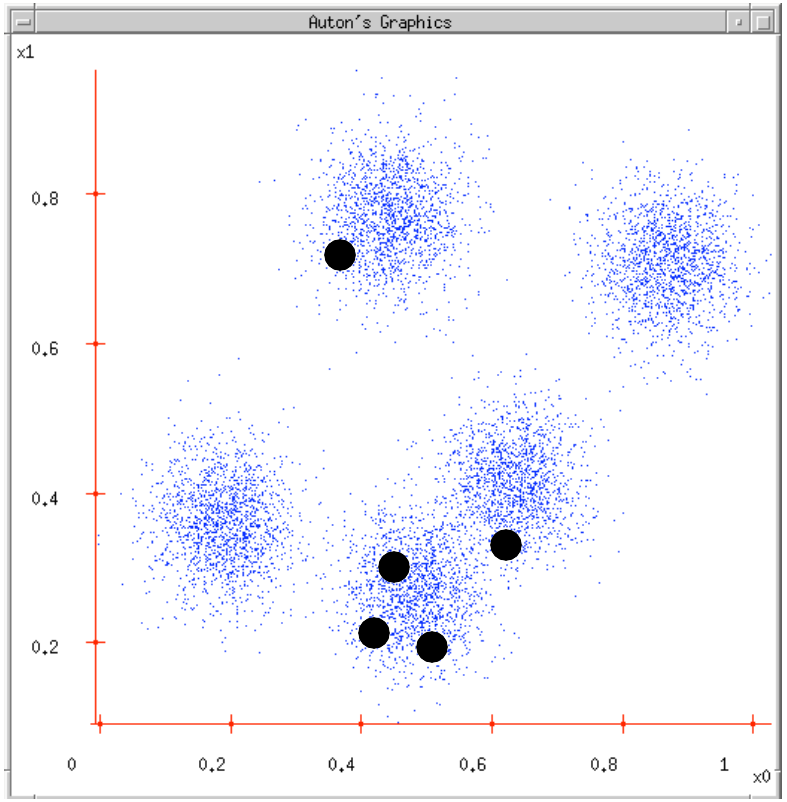
Clustering Data



K–Means Clustering

K–Means (k , X)

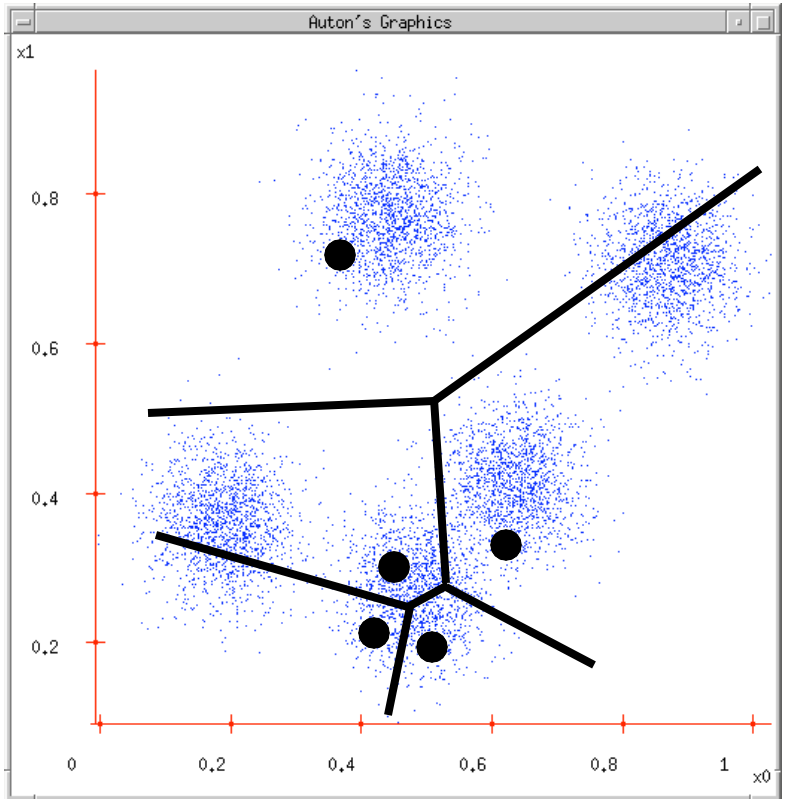
- Randomly choose k cluster center locations (centroids)
- Loop until convergence
  - Assign each point to the cluster of the closest centroid
  - Re–estimate the cluster centroids based on the data assigned to each cluster



K–Means Clustering

K–Means (k , X)

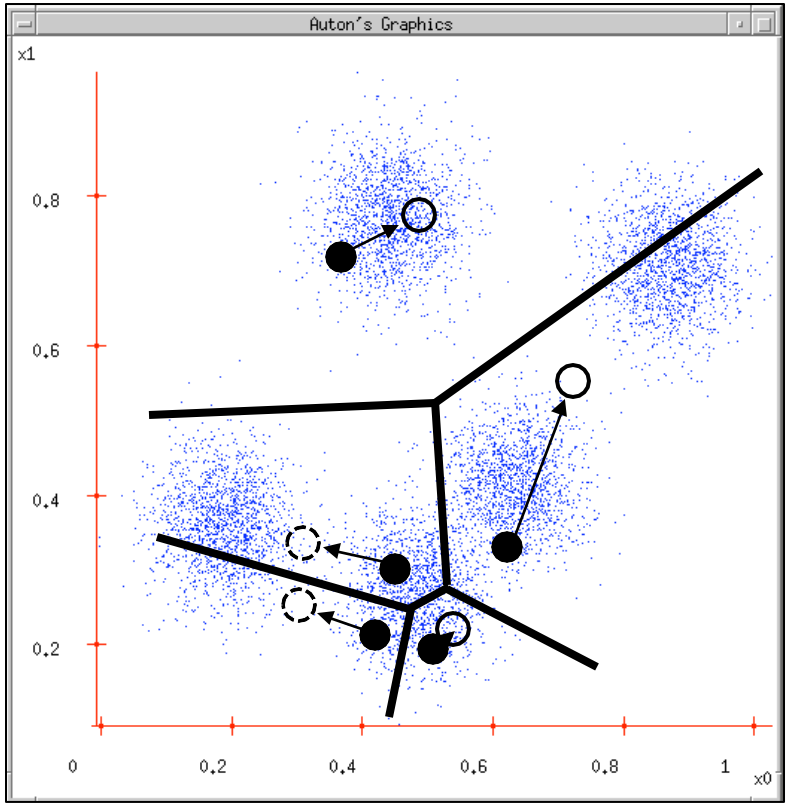
- Randomly choose k cluster center locations (centroids)
- Loop until convergence
  - Assign each point to the cluster of the closest centroid
  - Re–estimate the cluster centroids based on the data assigned to each cluster



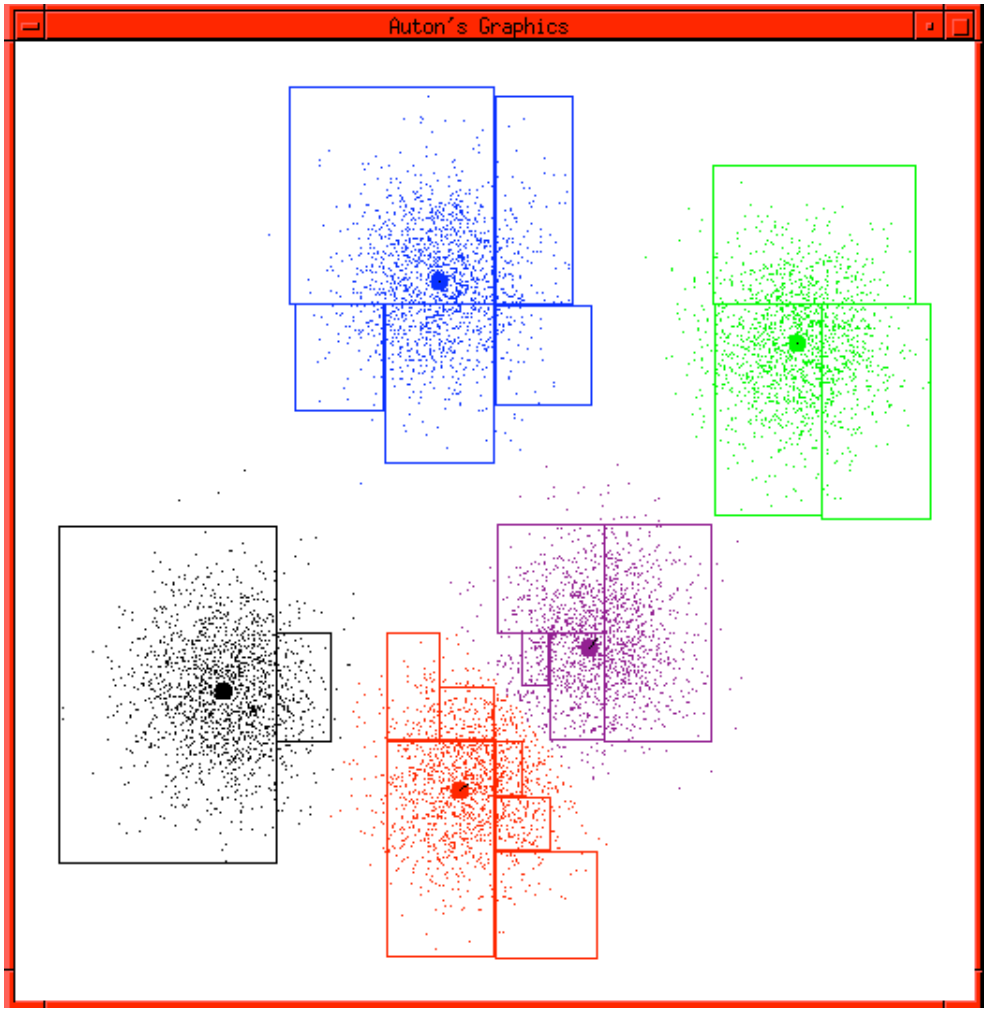
K–Means Clustering

K–Means (k , X)

- Randomly choose k cluster center locations (centroids)
- Loop until convergence
  - Assign each point to the cluster of the closest centroid
  - Re–estimate the cluster centroids based on the data assigned to each cluster



K–Means



Example generated by Andrew Moore using Dan Pelleg's super-duper fast K-means system:

Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999.

Visualizing K-Means

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K–Means Objective Function

- K–means finds a local optimum of the following objective function:

$$\arg\min_{\boldsymbol{S}} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \boldsymbol{\mu}_i\|_2^2$$

where $\boldsymbol{S} = \{S_1, \ldots, S_k\}$ is a partitioning over $X = \{x_1, \ldots, x_n\}$ s.t. $X = \bigcup_{i=1}^{k} S_i$

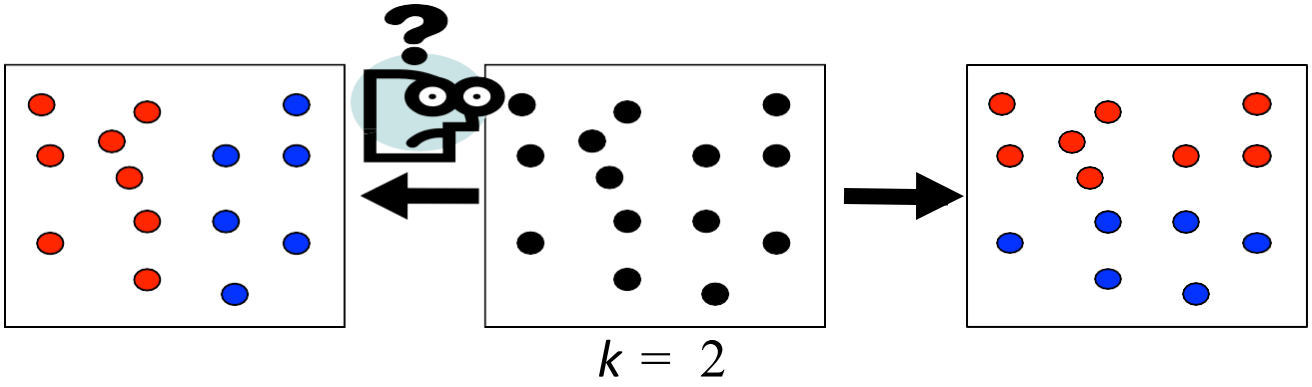
and $\boldsymbol{\mu}_i = \text{mean}(S_i)$

Problems with K–Means

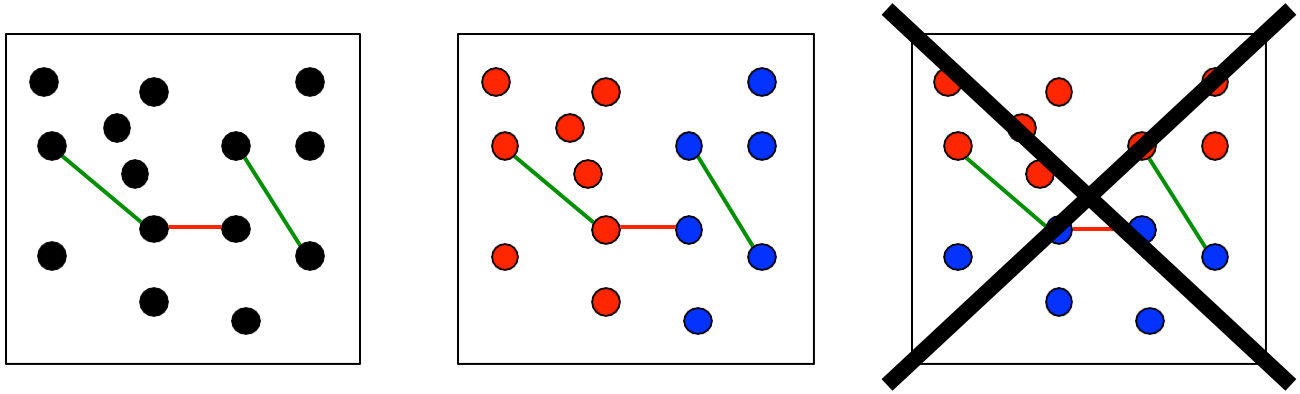
- ***Very*** sensitive to the initial points
  - Do many runs of K–Means, each with different initial centroids
  - Seed the centroids using a better method than randomly choosing the centroids
    - e.g., Farthest--first sampling
- Must manually choose *k*
  - Learn the optimal *k* for the clustering
    - Note that this requires a performance measure

Problems with K–Means

- How do you tell it which clustering you want?



Constrained clustering techniques (semi--supervised)



Same--cluster constraint (must--link)

Different--cluster constraint (cannot--link)

Generative Modeling

Discriminative vs Generative Modeling

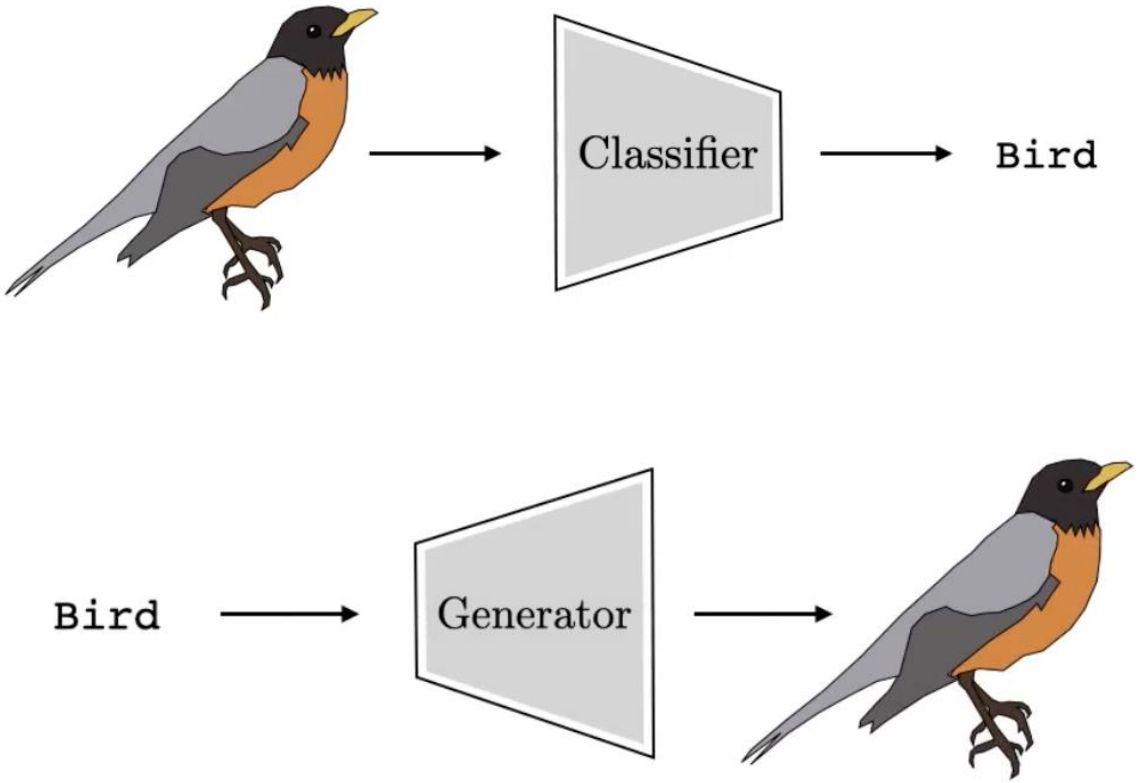
- **Discriminative modeling** focuses on learning the decision boundary between classes.
  - Common in **classification** tasks, e.g., logistic regression, neural networks.

- **Generative modeling** learns the **data distribution** (e.g., p(x) or p(x, y))
  - Can be used to **generate** new data similar to the training set.

- **Important:** Not all generative models are unsupervised!
  - Some, like **Naive Bayes** are trained with labels.

Discrimination vs. Generation

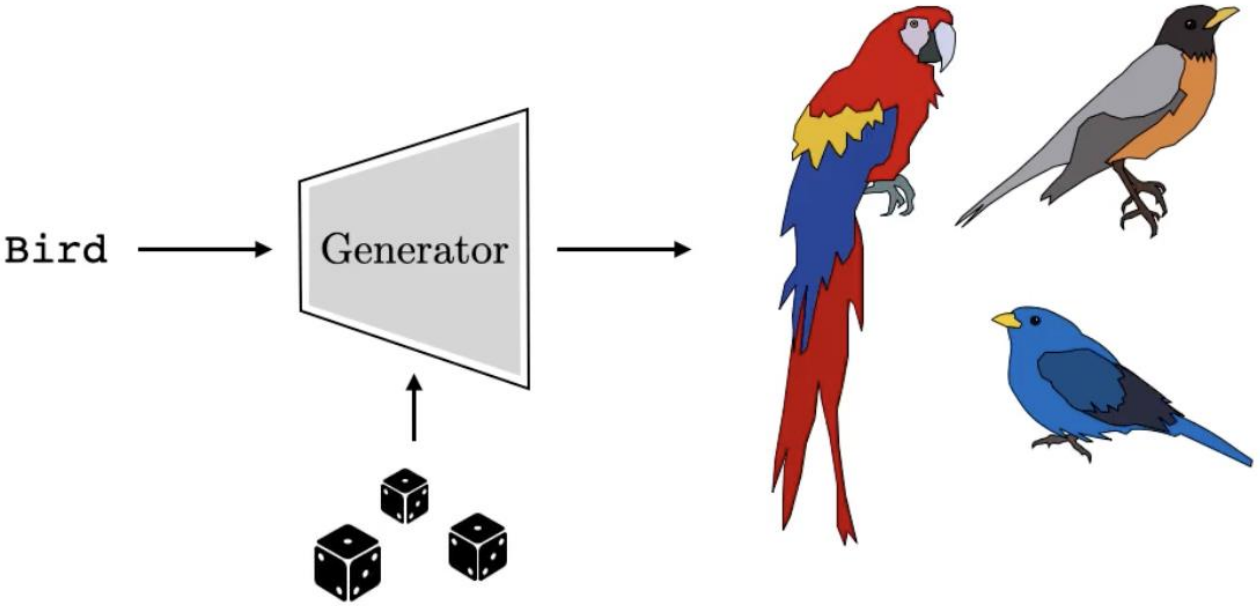
- Generation is the “inverse” process



Slide Credit: Angie Liu

Generative Models

- A generative model takes in a **random noise vector** (like rolling dice), and produces a **realistic-looking sample** from the learned data distribution.



Some Generative Models

- VAE (Variational Autoencoder)
- GANs (Generative Adversarial Networks)
- Diffusion Models