Quiz 10 (on Canvas)

Ends at 1:06pm

CS5670: Computer Vision

Diffusion models

“A copy of a computer vision textbook entitled ‘Szeliski 2nd Edition’ sitting on a beautiful coffee table”
(according to ChatGPT 4)

Announcements

- In class final this coming Tuesday, May 7
  - 2 sheets of notes (front and back) allowed
  - Final is comprehensive (covers entire course)
- Course evaluations are open
  - We would love your feedback!
  - Small amount of extra credit for filling out
    - What you write is still anonymous; instructors only see if students filled it out
  - https://apps.engineering.cornell.edu/CourseEval/

Readings

- 5-Minute Graphics from Steve Seitz:
  - Large Language Models from scratch
  - Large Language Models: Part 2
  - Text to Image in 5 minutes: Parti, Dall-E 2, Imagen
  - Text to Image: Part 2 -- how image diffusion works in 5 minutes

Recall: The Space of All Images

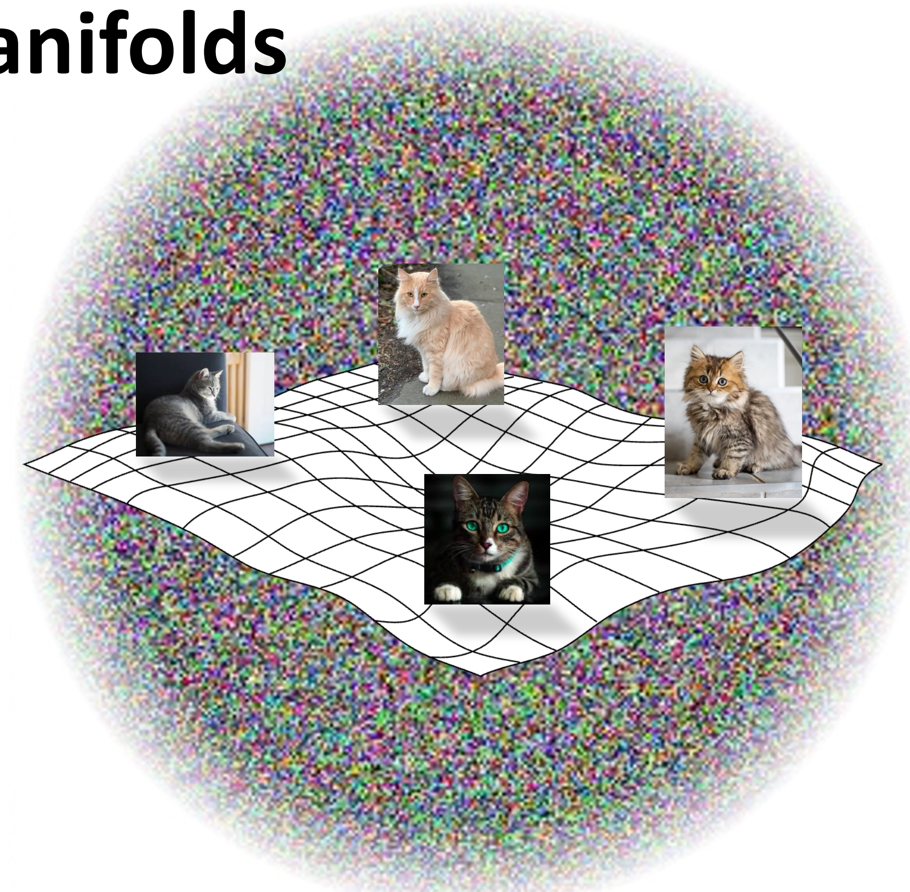
- Lets consider the space of all 100x100 images
- Now lets randomly sample that space…
- Conclusion: Most images are noise

Question:
What do we expect a random uniform sample of all images to look like?

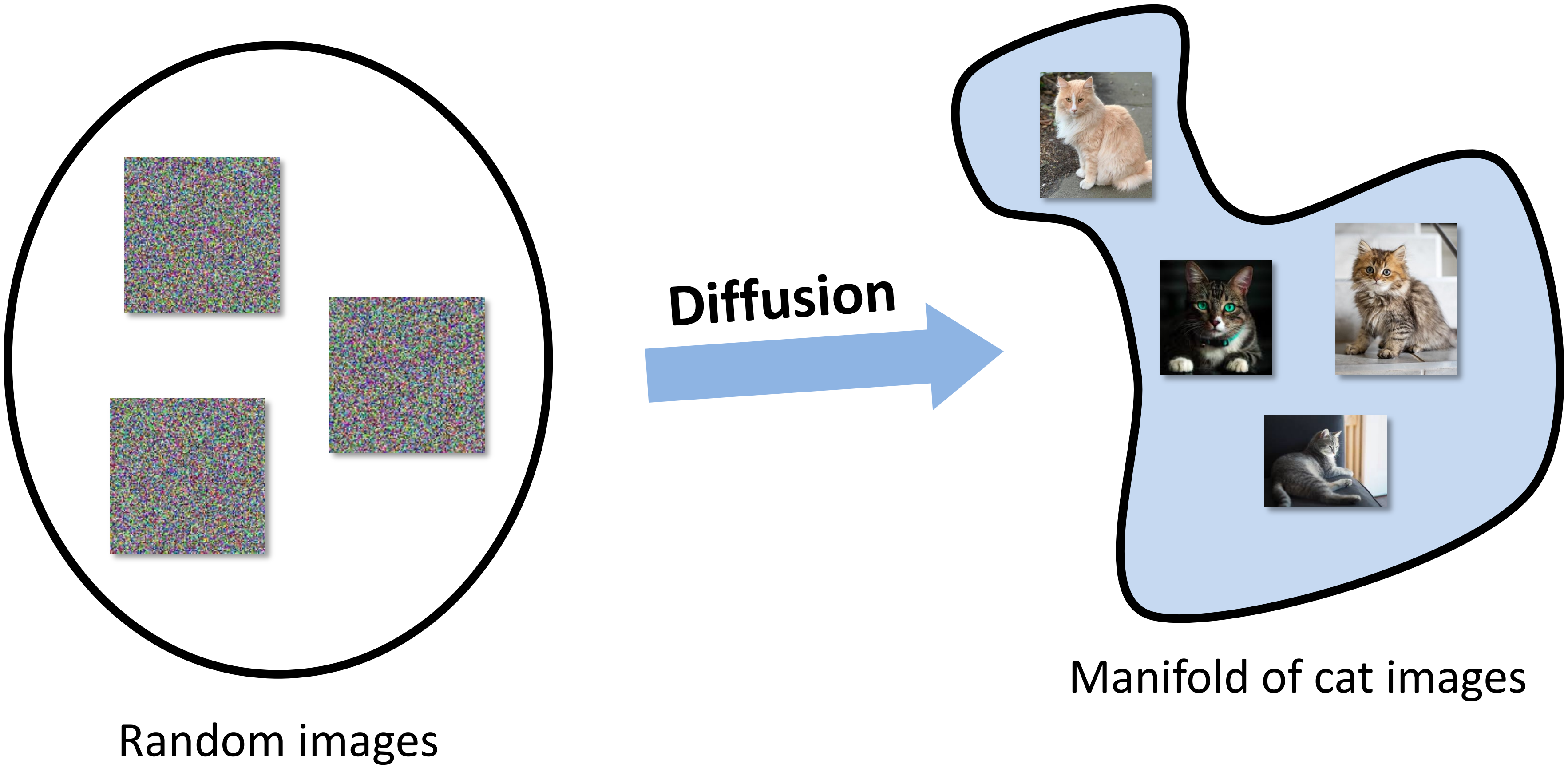
```
pixels = np.random.rand(100,100,3)
```

Recall: Natural Image Manifolds

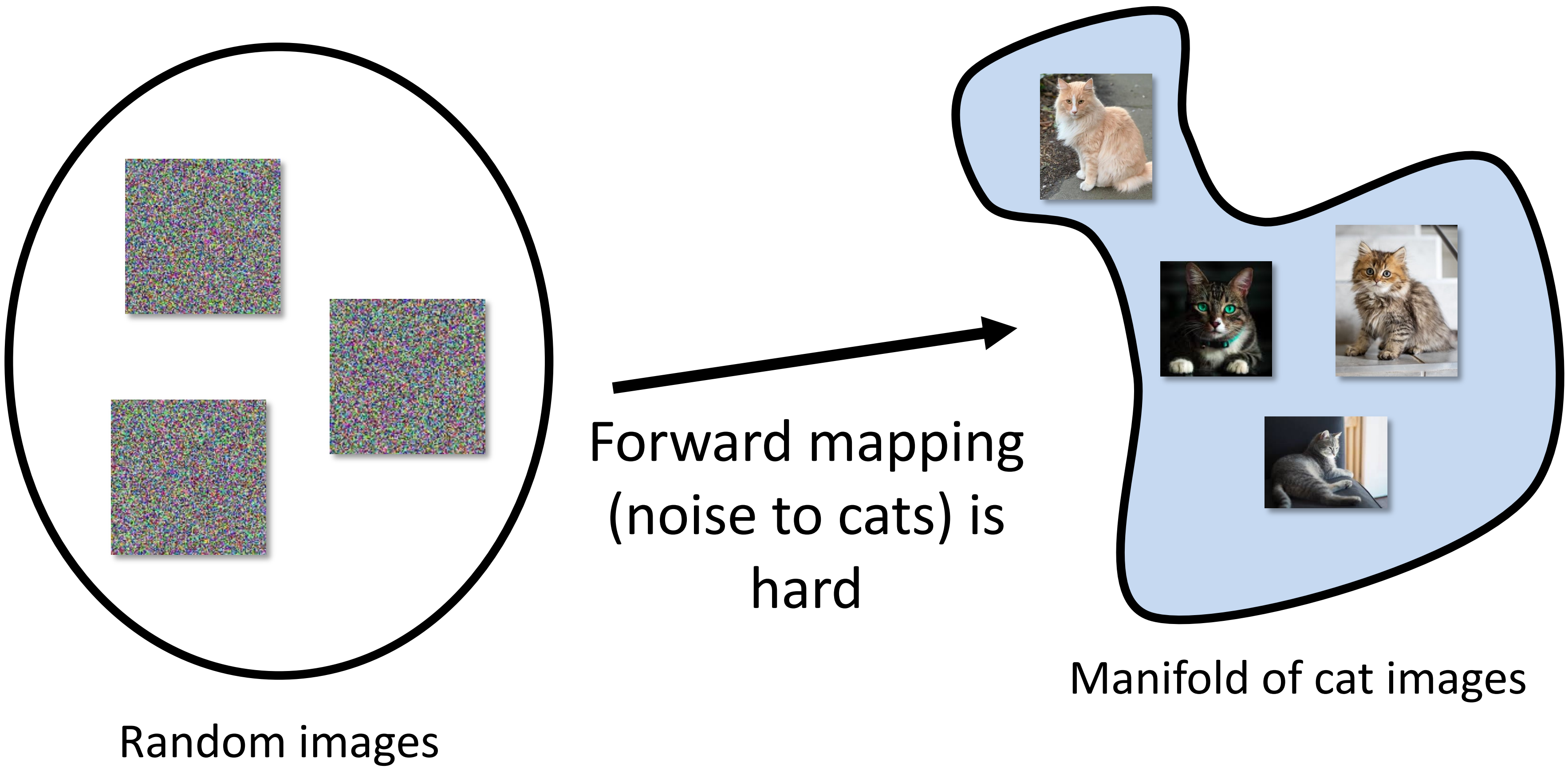
- Most images are "noise"
- "Meaningful" images tend to form some manifold within the space of all images
- Images of a particular class fall on manifolds within that manifold…



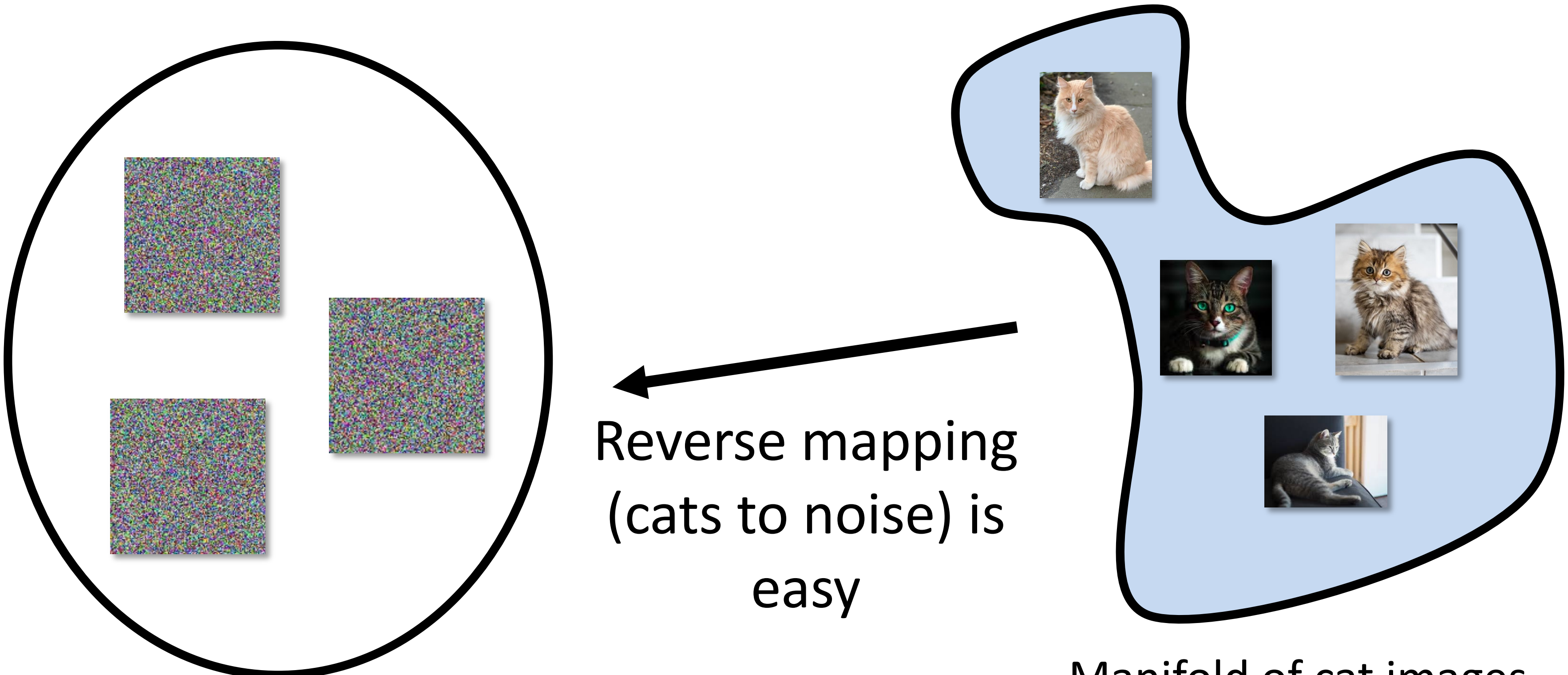
The Space of All Images



Slide concept: Steve Seitz



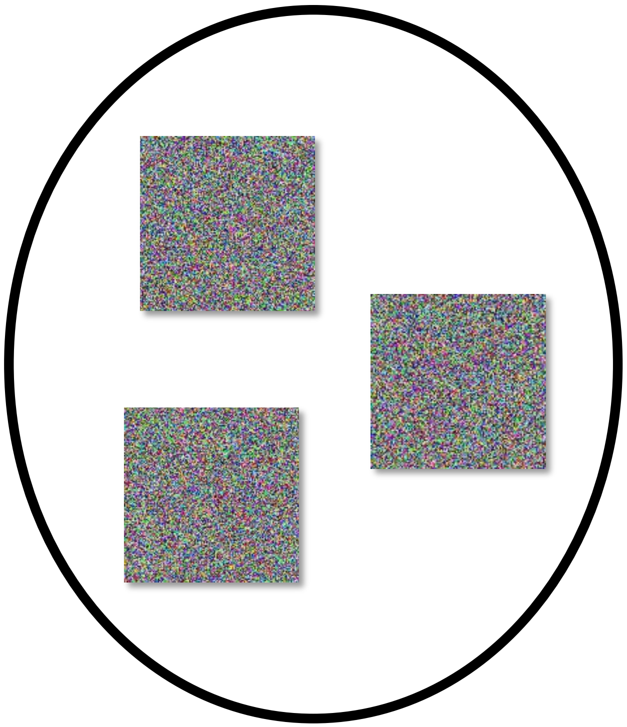
Slide concept: Steve Seitz



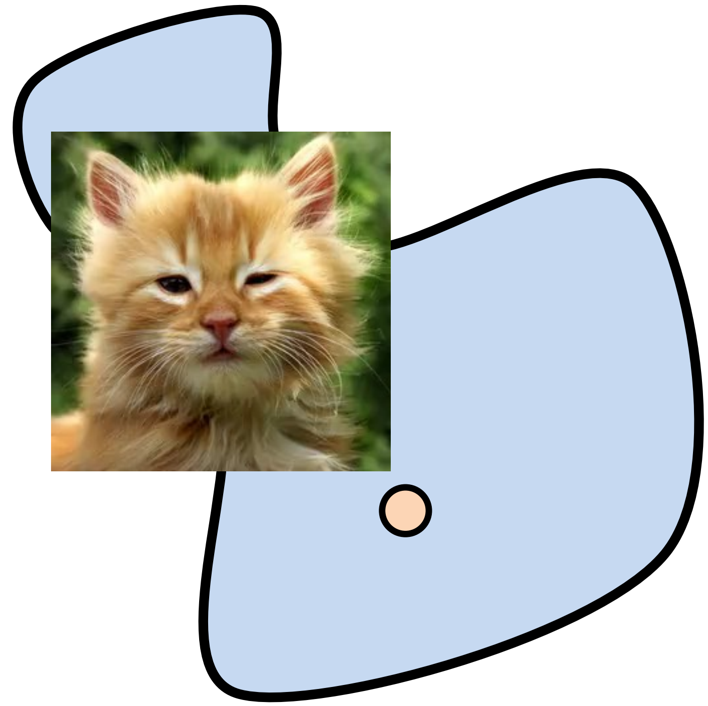
Random images

Manifold of cat images

Slide concept: Steve Seitz

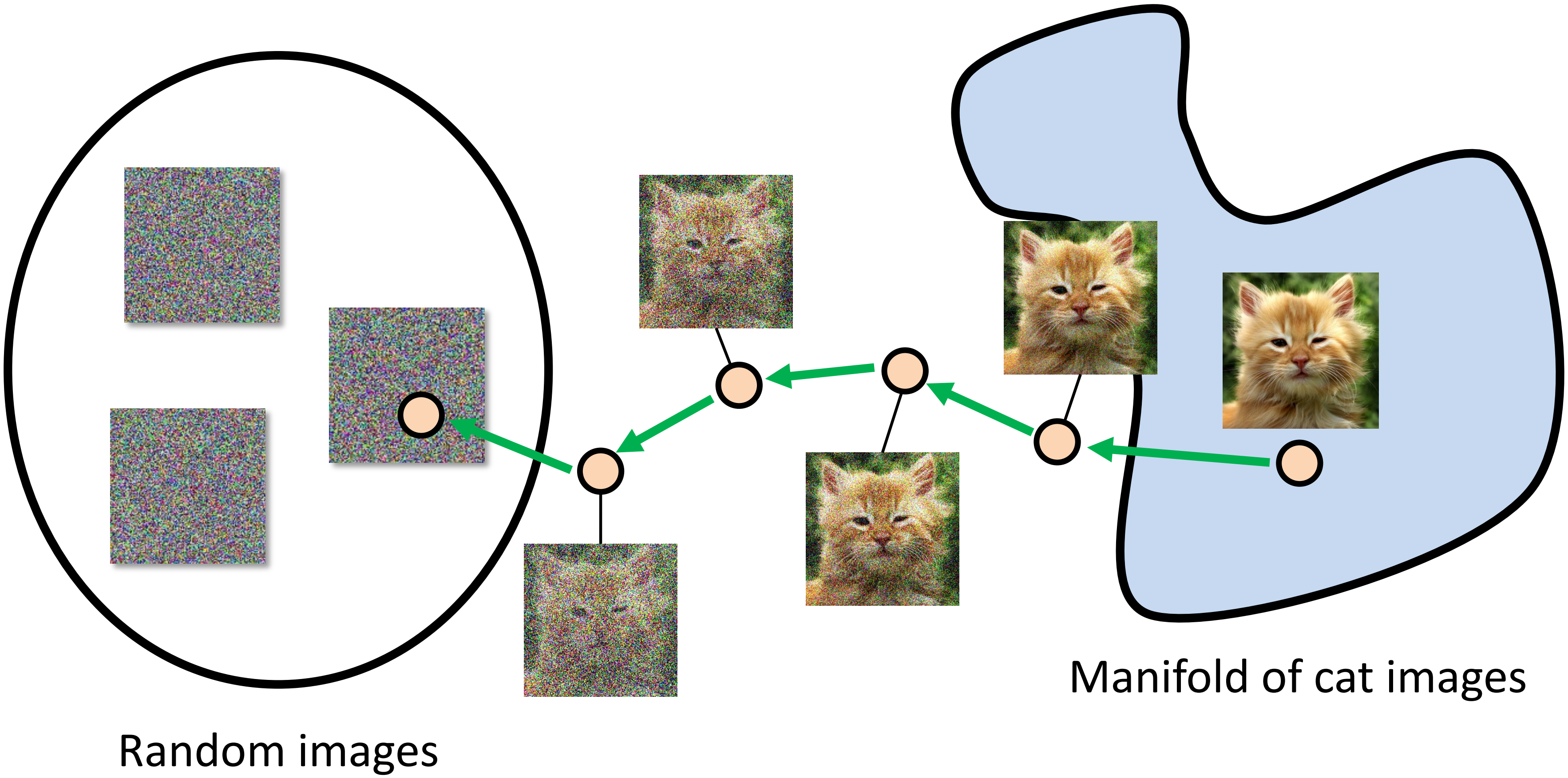


Random images

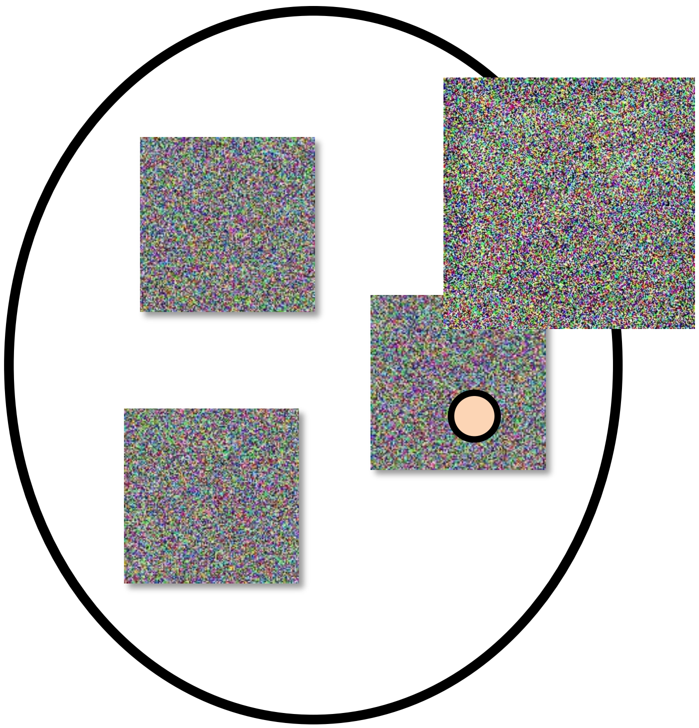


Manifold of cat images

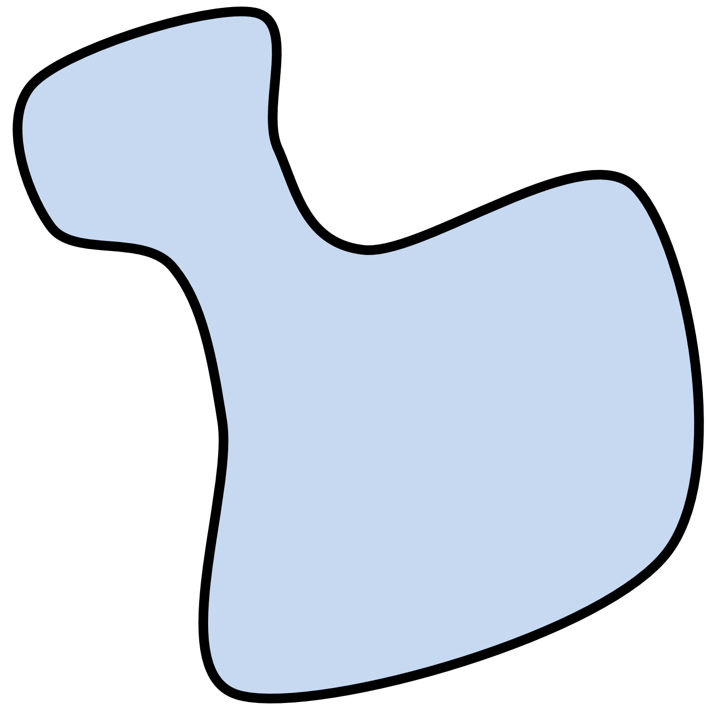
Slide concept: Steve Seitz



Slide concept: Steve Seitz

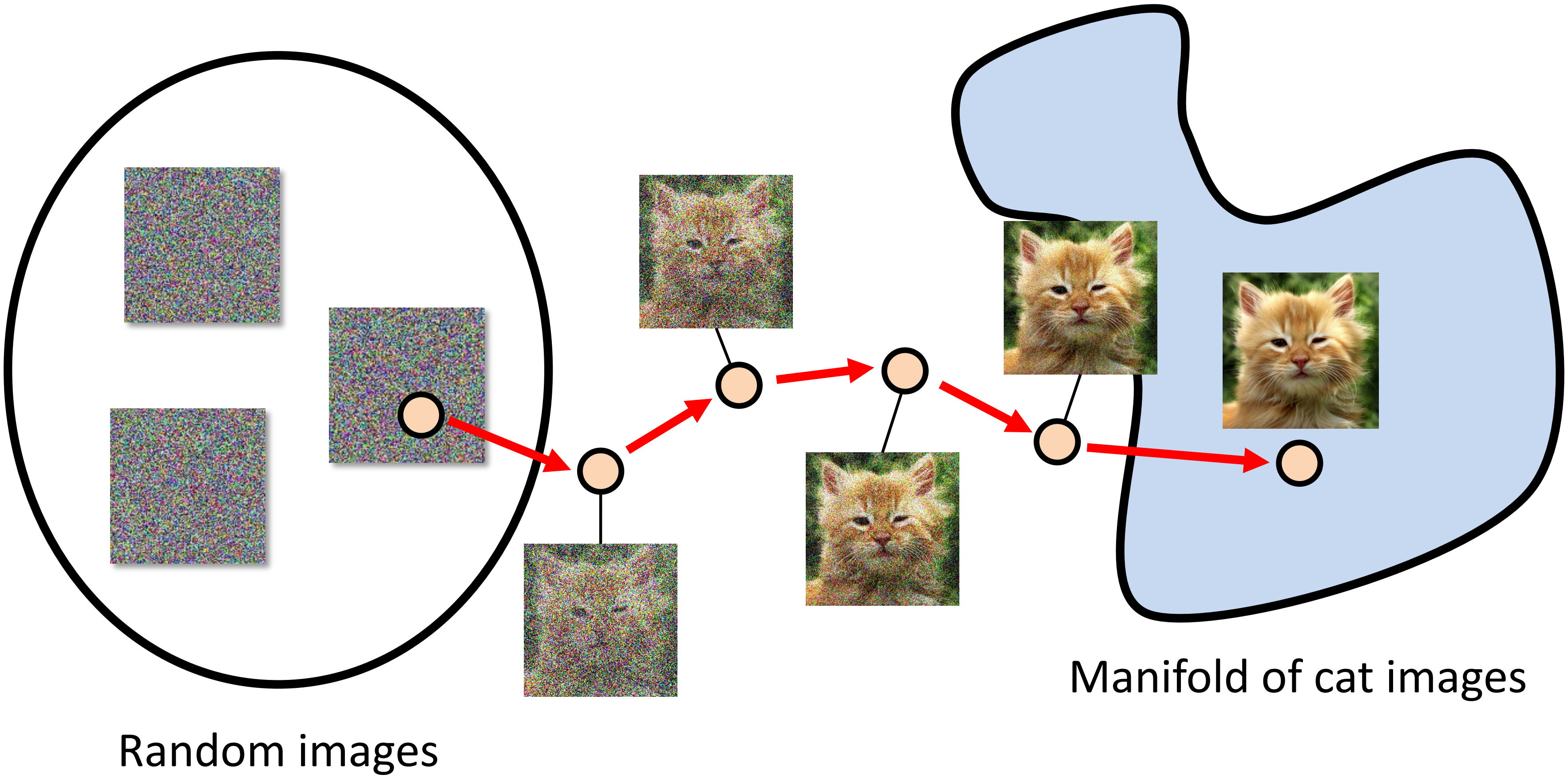


Random images

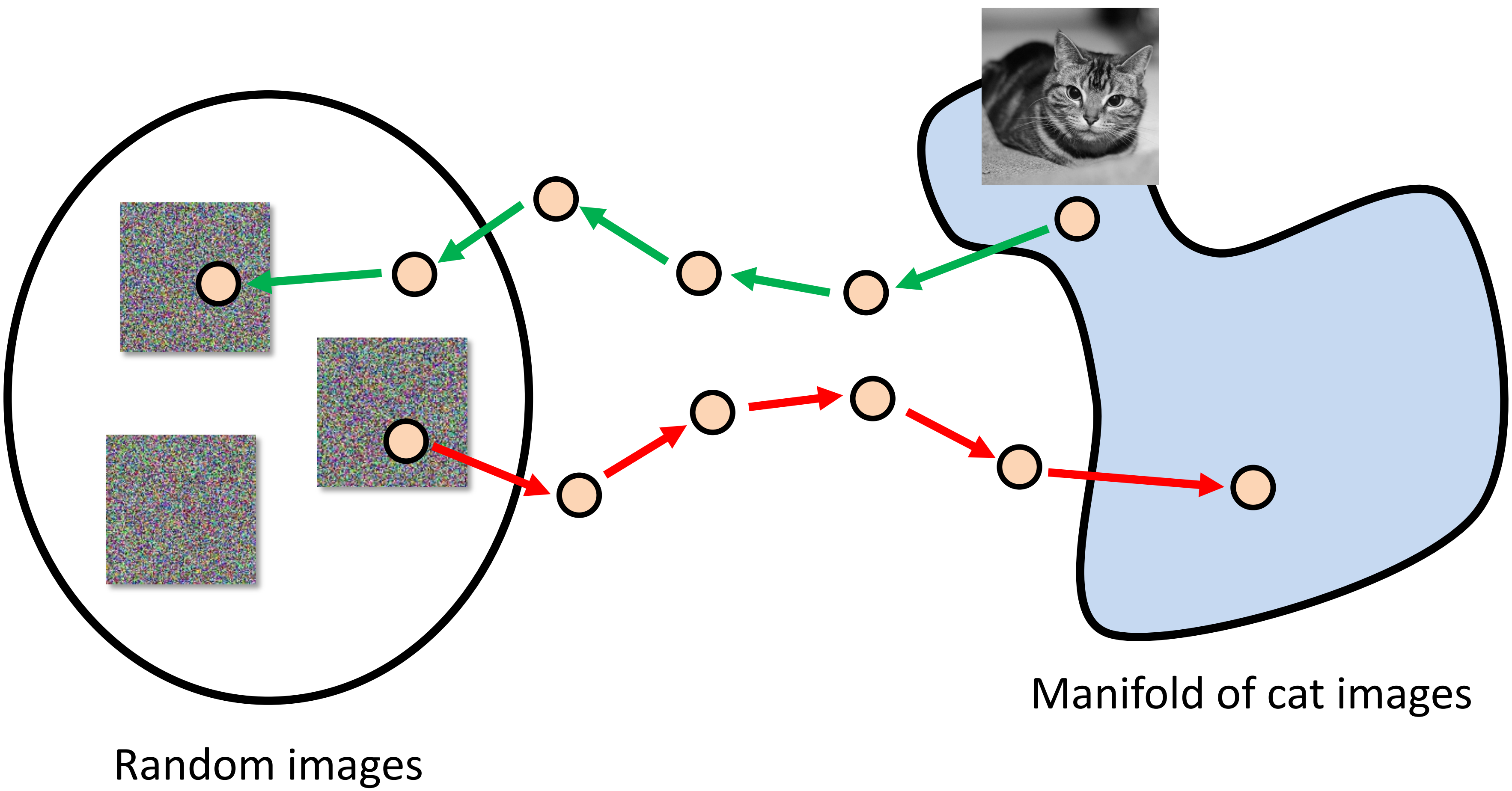


Manifold of cat images

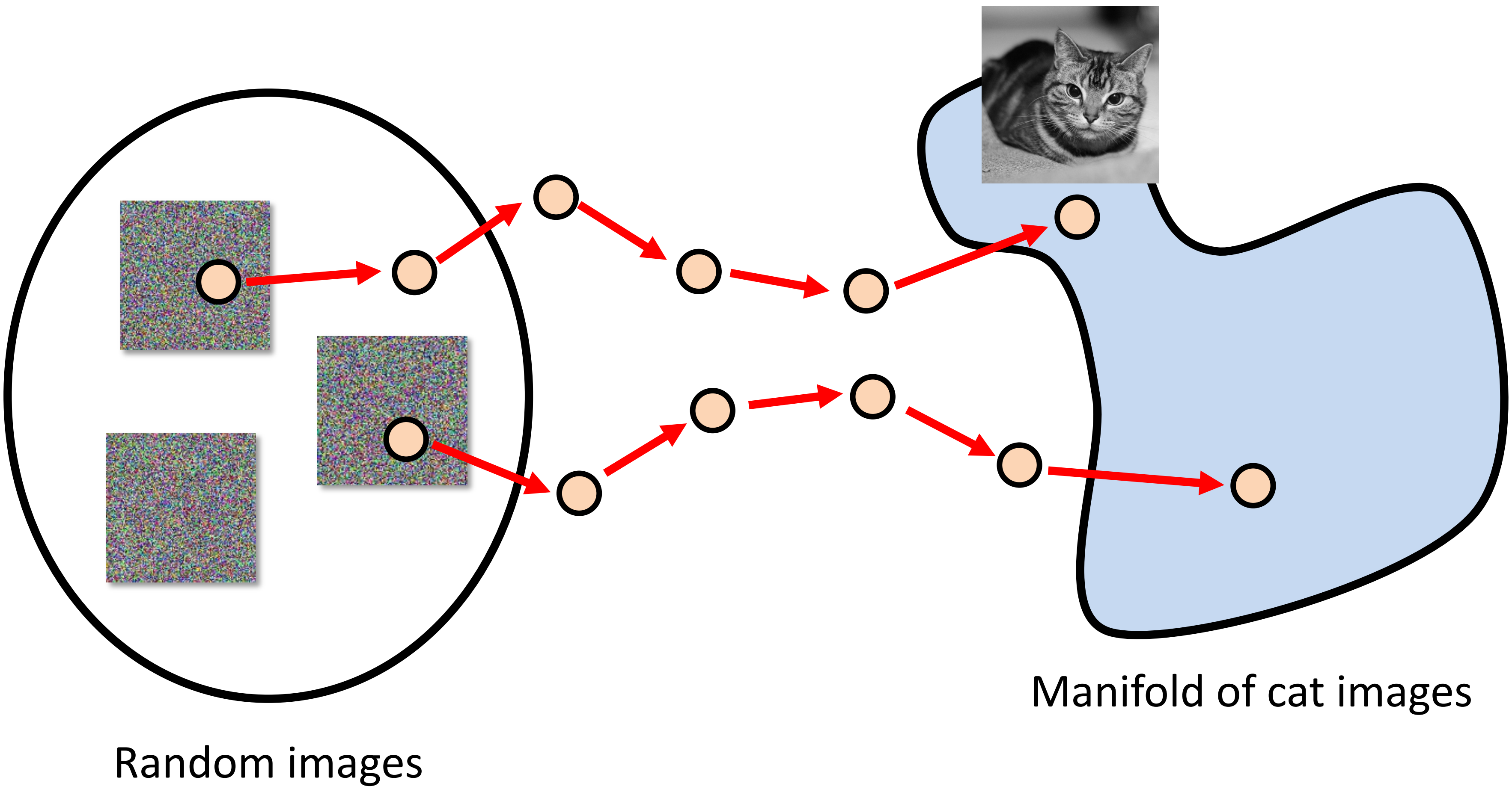
Slide concept: Steve Seitz



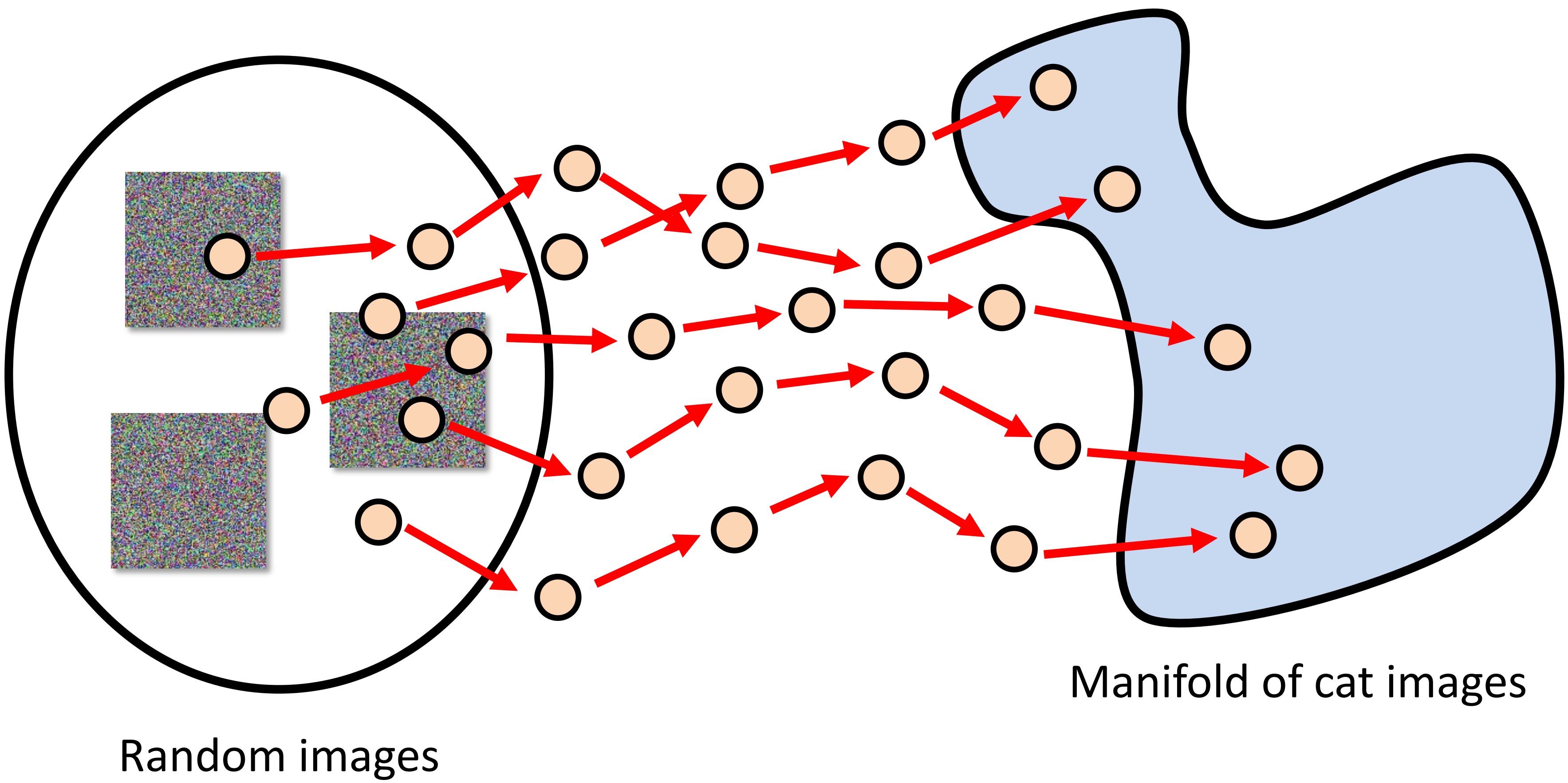
Slide concept: Steve Seitz



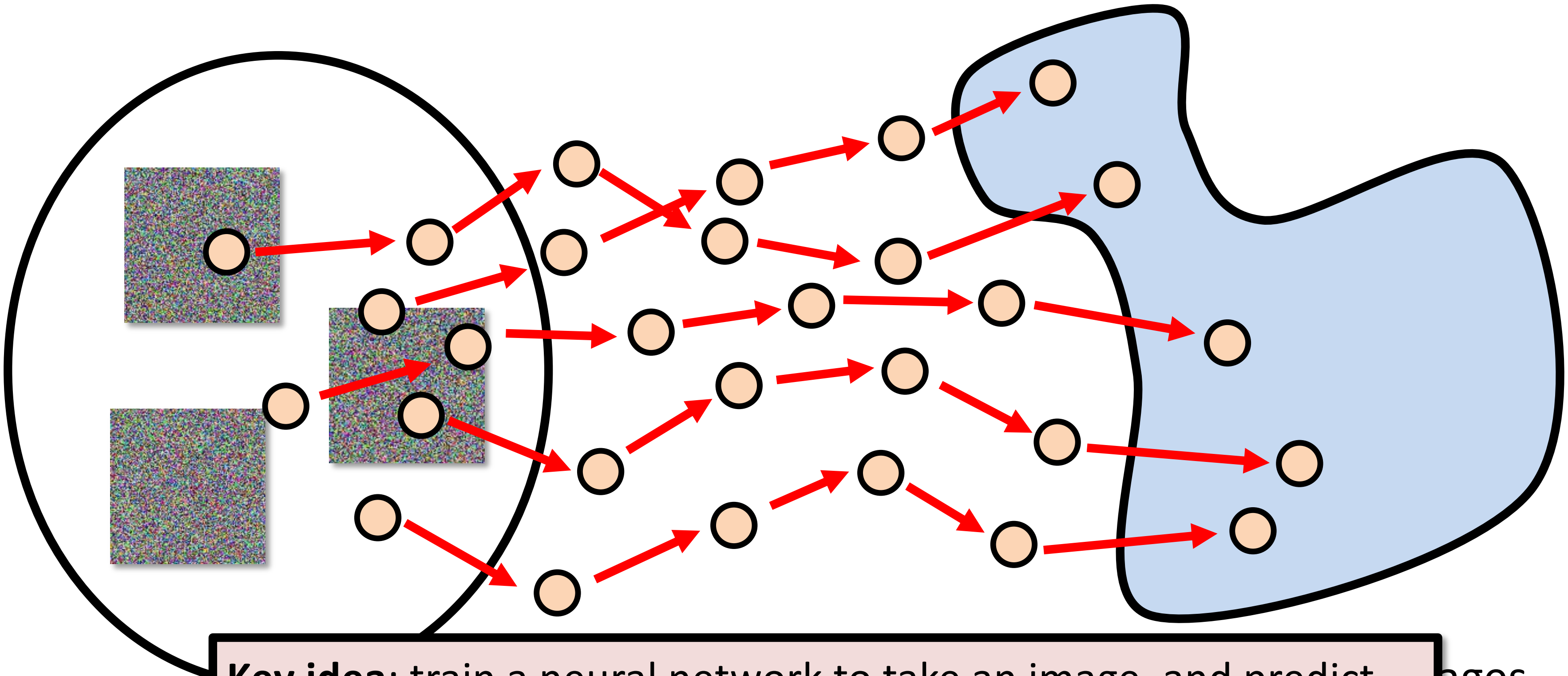
Slide concept: Steve Seitz



Slide concept: Steve Seitz



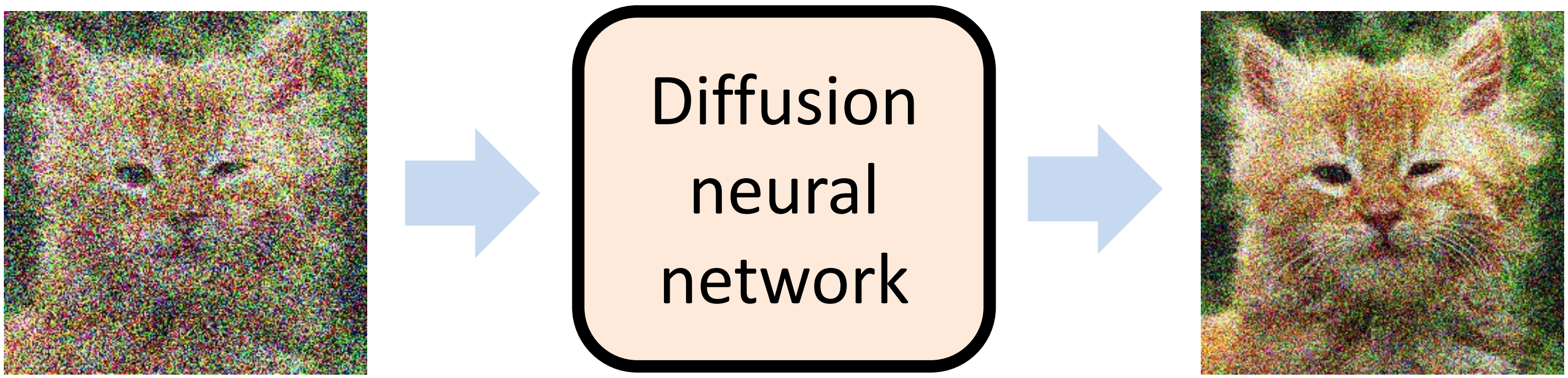
Slide concept: Steve Seitz



Key idea: train a neural network to take an image, and predict the corresponding arrow above; that is, predict to convert a noisy image to a slightly less noisy image that is closer to the desired image manifold, using the examples above to train.

Slide concept: Steve Seitz

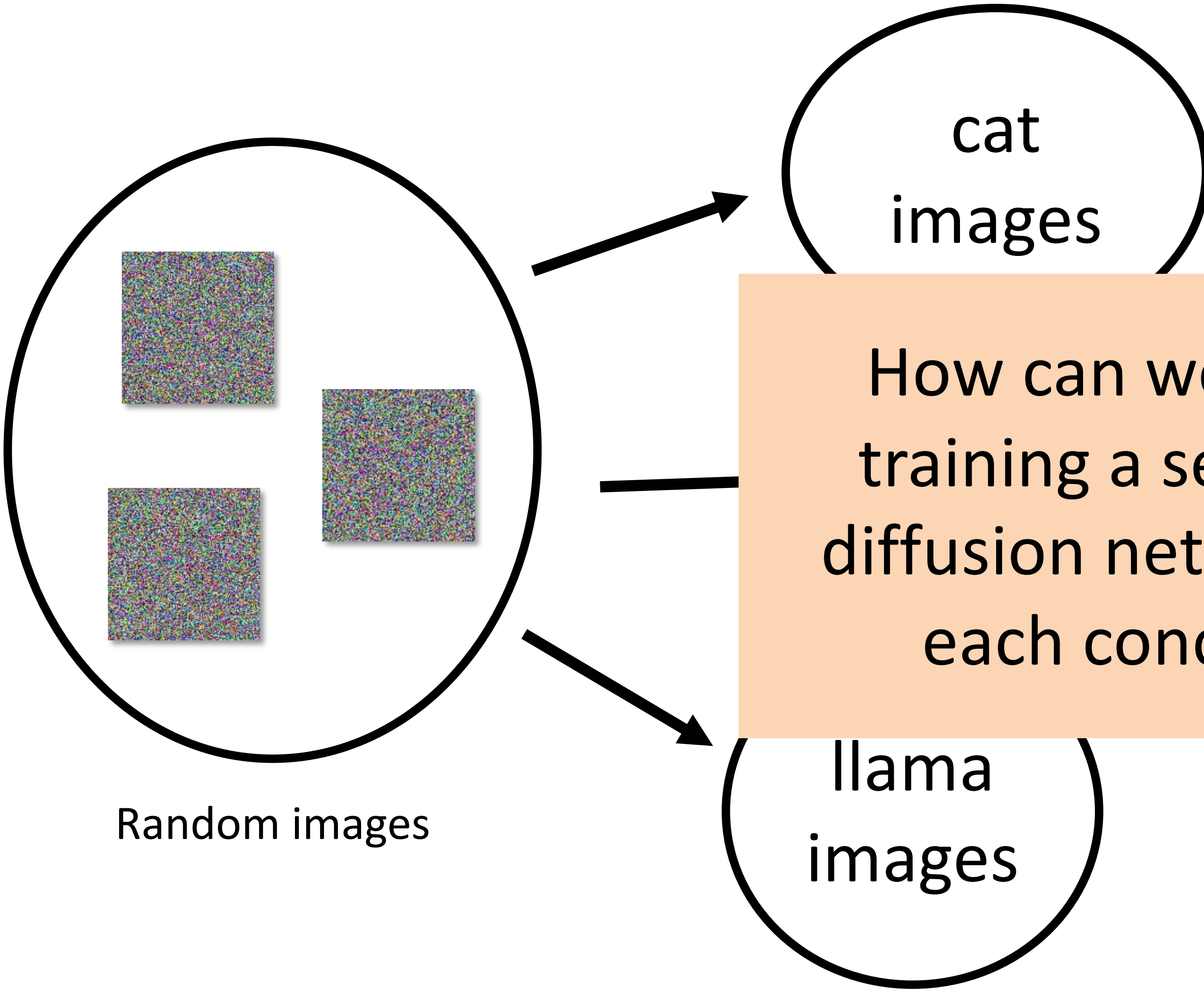
Denoising diffusion neural network



This network can be a U-Net or other suitable image-to-image network

Generating new images

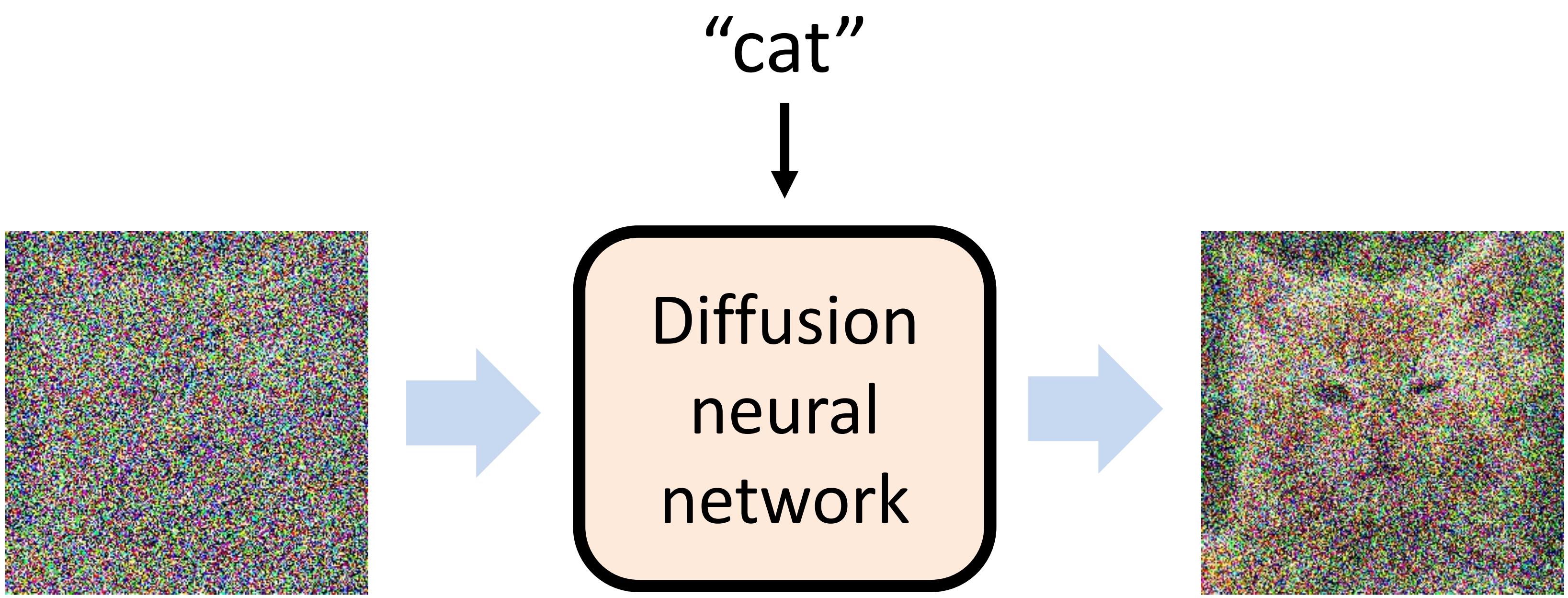
- Once diffusion network has been trained, generate new images by starting with a random noise image, and iteratively applying the network to slowly remove noise, for some number of steps (e.g., 1,000 for DALL-E 2)
- "Walking from random images towards the manifold of natural images"



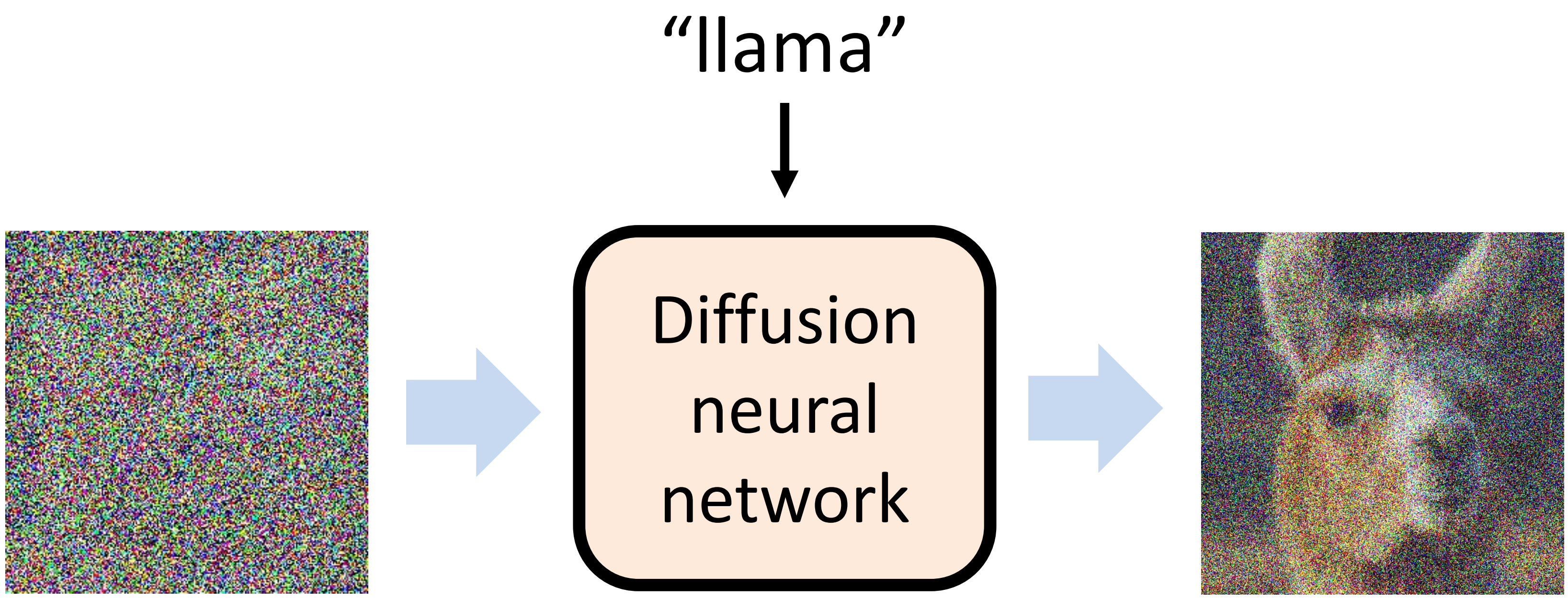
How can we avoid training a separate diffusion network for each concept?

Slide concept: Steve Seitz

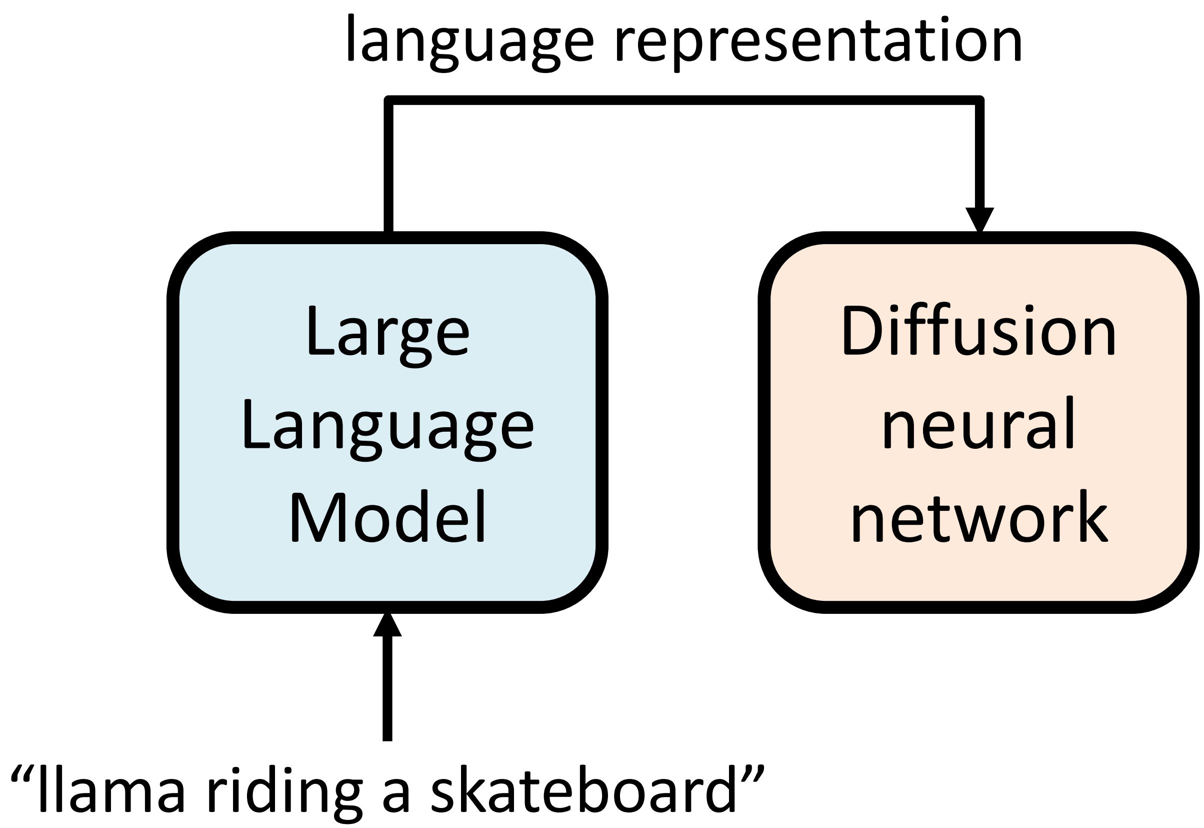
Idea 1: add a text label as conditioning



Idea 1: add a text label as conditioning



Idea 2: condition using large language model



Training on images + captions

A pack llama in the Rocky Mountain National Park

https://en.wikipedia.org/wiki/Llama

DALL-E 2

"A llama riding a skateboard"

"A llama riding a skateboard captured with a DSLR"

Imagen

“Sprouts in the shape of text 'Imagen' coming out of a fairytale book.”

“A dragon fruit wearing karate belt in the snow.”

Other applications of diffusion models

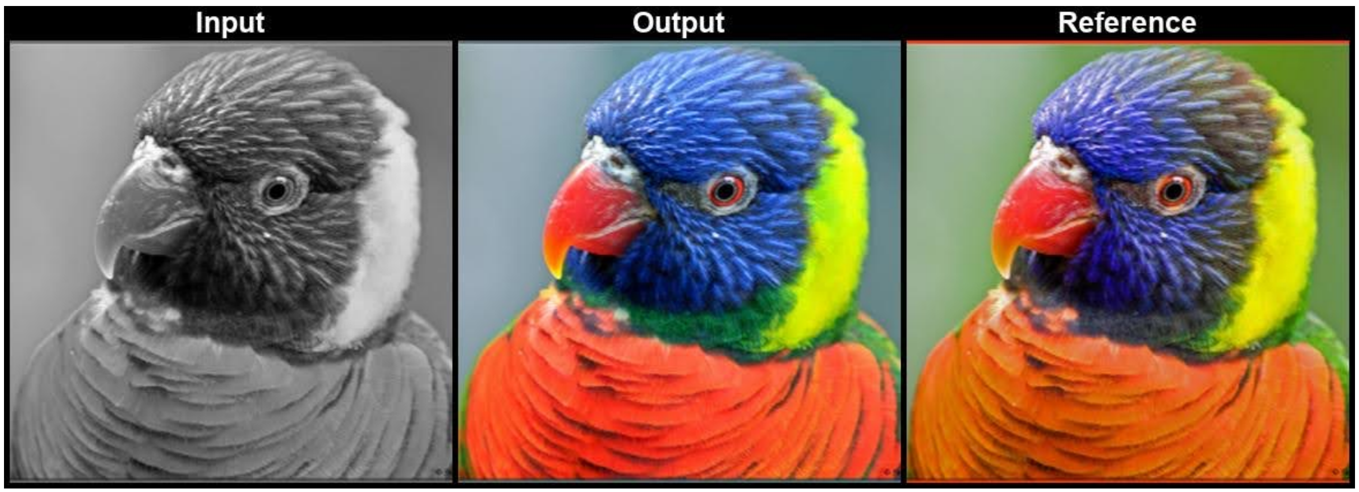
- Uncropping

Progressively zooming out. The most zoomed-in image is the input

Palette: Image-to-Image Diffusion Models
Saharia et al. arXiv 2022.

Other applications of diffusion models

- Colorization



[Palette: Image-to-Image Diffusion Models]
Saharia et al. arXiv 2022.

Other applications of diffusion models

- Inpainting

[Palette: Image-to-Image Diffusion Models](#)
Saharia et al. arXiv 2022.

DreamFusion: Text-to-3D using 2D Diffusion

“a DSLR photo of a squirrel”

https://dreamfusion3d.github.io/

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz Yuanzhen Li Varun Jampani Yael Pritch Michael Rubinstein Kfir Aberman

Google Research

It's like a photo booth, but once the subject is captured, it can be synthesized wherever your dreams take you…

[Paper] (new!) [Dataset] [BibTeX]

Comparison with GANs

- Diffusion models tend to be easier to train and more scalable
- Diffusion models tend to be slower – often many iterations of denoising are required
- However, recent work is mitigating some of these issues (with both GANs and diffusion models)

Text-to-image model zoo

- Diffusion models
  - DALL-E 2/3, Imagen, Stable Diffusion
- Transformer-based models
  - DALL-E, Parti, MUSE

Questions?