CausalGAN
Learning Causal Implicit Generative Models with Adversarial Training

Murat Kocaoglu
MIT-IBM Watson AI Lab
Cambridge, MA

WHY’19 Symposium
March 25, 2019

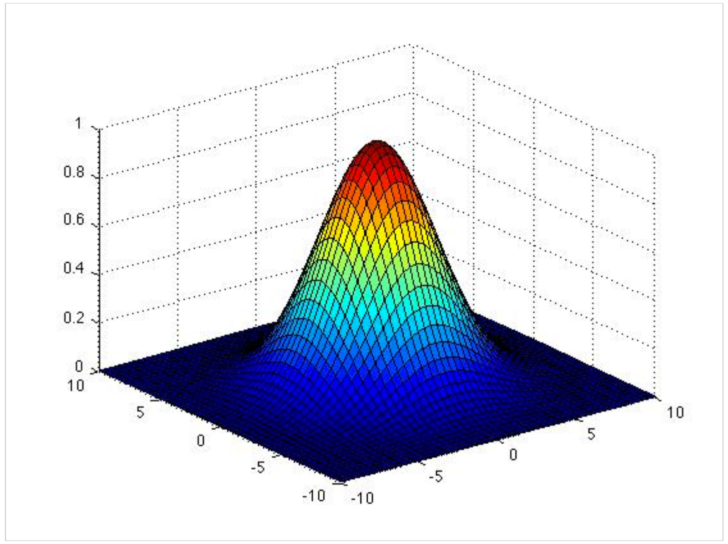
Based on joint work with

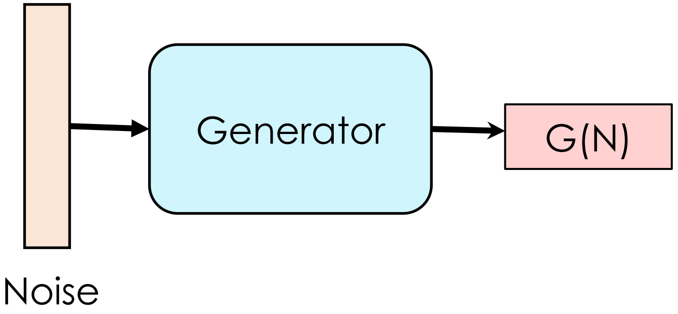
Chris Snyder
Alex Dimakis
Sriram Vishwanath

Implicit Generative Models

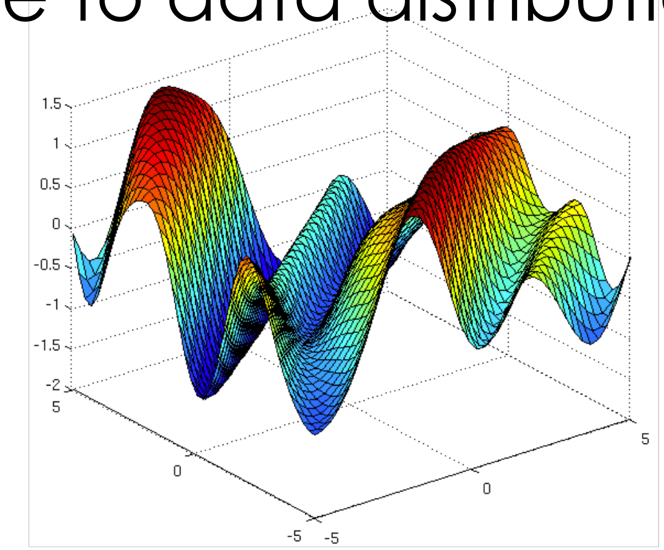
- Allow sampling from distribution without explicit parameterization.

- Learn a mapping from known noise to data distribution.



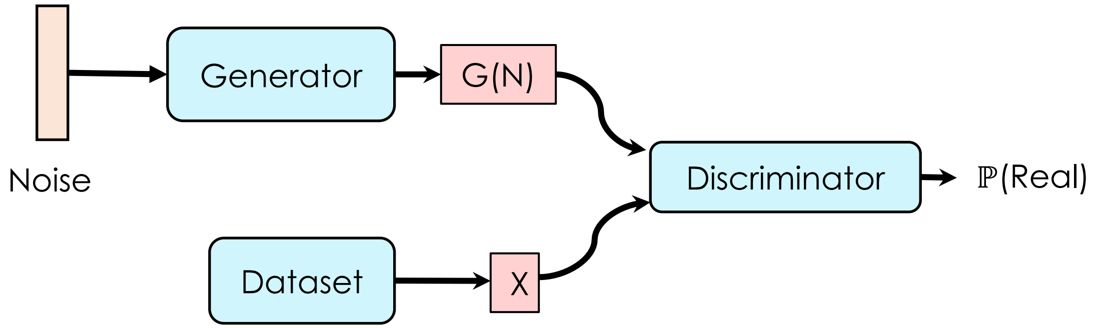


$G(N) \sim p_{data}$



Generative Adversarial Networks
[Goodfellow'14]

- Learning distribution with the help of adversary



- Generator – discriminator optimize opposite objectives

$$\min_{G}\max_{D}\mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))]+\mathbb{E}_{x\sim p_g(x)}[\log(1-D(x))]$$

- Iterative training with stochastic gradient descent

Improved GANs

- WGAN[Arjovsky'17]:
  Optimizing Wasserstein metric instead of Jensen-Shannon divergence - more stable training

- BEGAN[Berthelot'17]:
  Use auto-encoder in the discriminator. More realistic face images

- Many more (ProgressiveGAN, StyleGAN)

  See
  https://github.com/hindupuravinash/the-gan-zoo

Weaknesses of GANs
[Other than actually training them]

- Can only sample from given data distribution.

- No way to “dream of” new distributions.

Weaknesses of GANs
[Other than actually training them]

- Can only sample from given data distribution.

- No way to "dream of" new distributions.

Our idea: Use causal knowledge to generate samples from interventional distributions.

Weaknesses of GANs
[Other than actually training them]

- Can only sample from given data distribution.

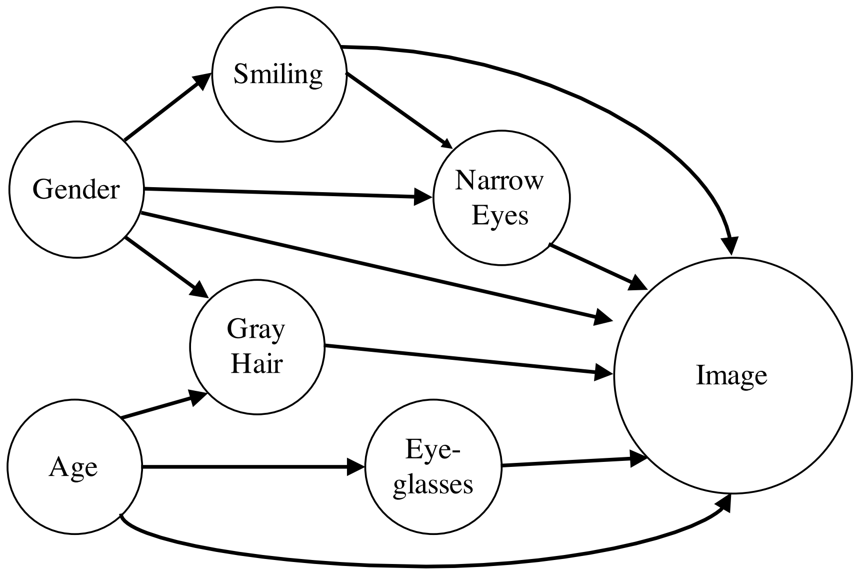
- No way to “dream of” new distributions.

Our idea: Use causal knowledge to generate samples from interventional distributions.

- Application:
  Causal image generation with labels.

Bringing Causality into Generative Models

- Image generation w/ labels as a causal process
- Assume causal graph is given
- Assume *Image* is always the sink node

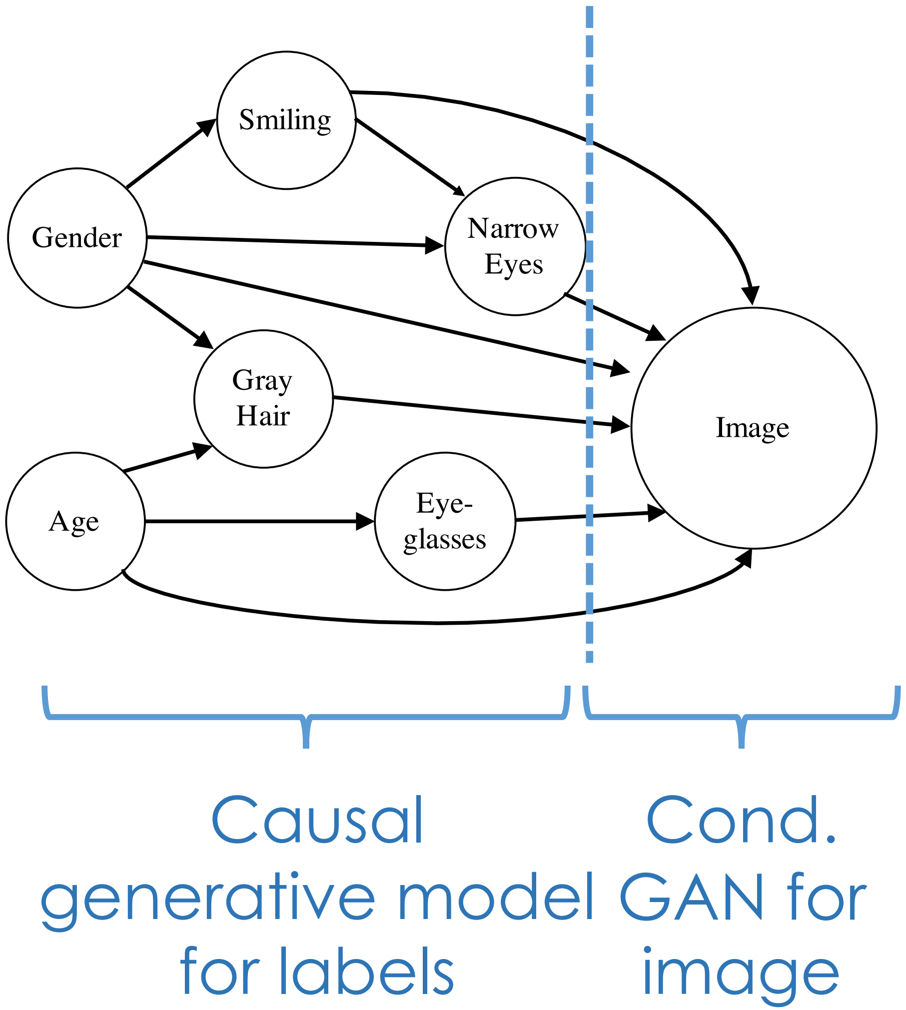


Challenge 1: Need to capture the causal structure.

Challenge 2: Training binary variables alongside image is difficult.

Bringing Causality into Generative Models

1. How to capture causal models with neural nets
2. Train causal generative model for labels
3. Train a conditional GAN to sample the image given labels
4. Combine label and image generation



Causal Models from Neural Nets

- Causal graph

$$X \to Z \leftarrow Y$$

- Structural equations:

$$X = f(E_X), \quad Y = g(E_Y), \quad Z = h(X, Y, E_Z)$$

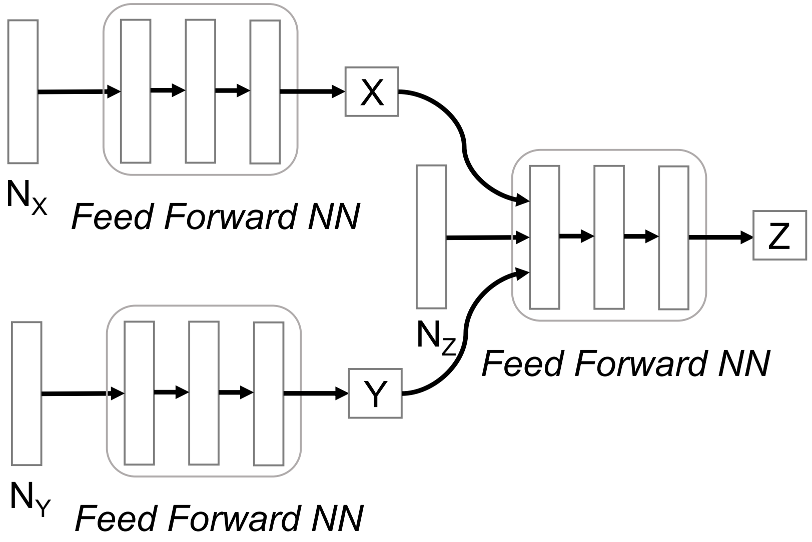
Causal Models from Neural Nets

- Causal graph

$$X \rightarrow Z \leftarrow Y$$

- Structural equations:

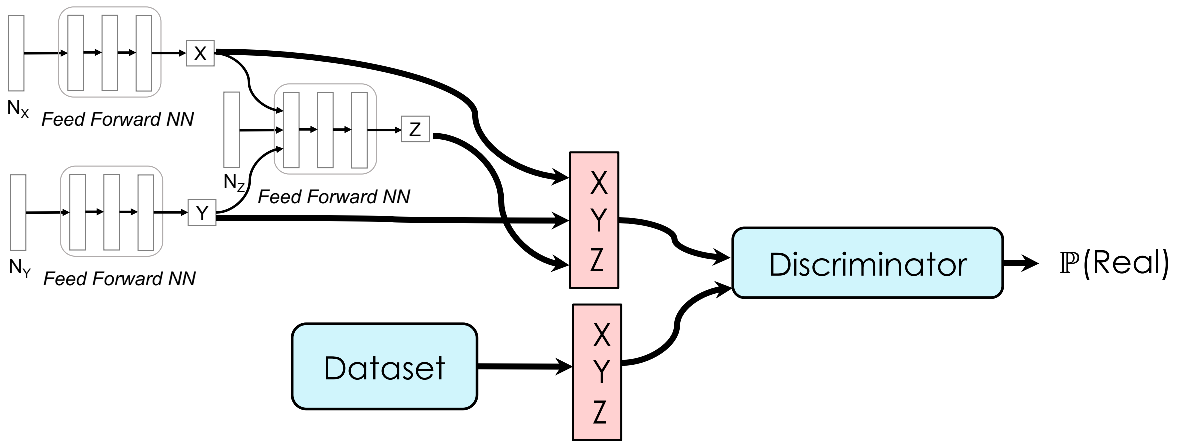
$$X = f(E_X), \quad Y = g(E_Y), \quad Z = h(X, Y, E_Z)$$



Training Causal Implicit Generative Models

- Structure the generator based on causal graph

- Use GAN training
  Remark: Wasserstein GAN training for discrete labels

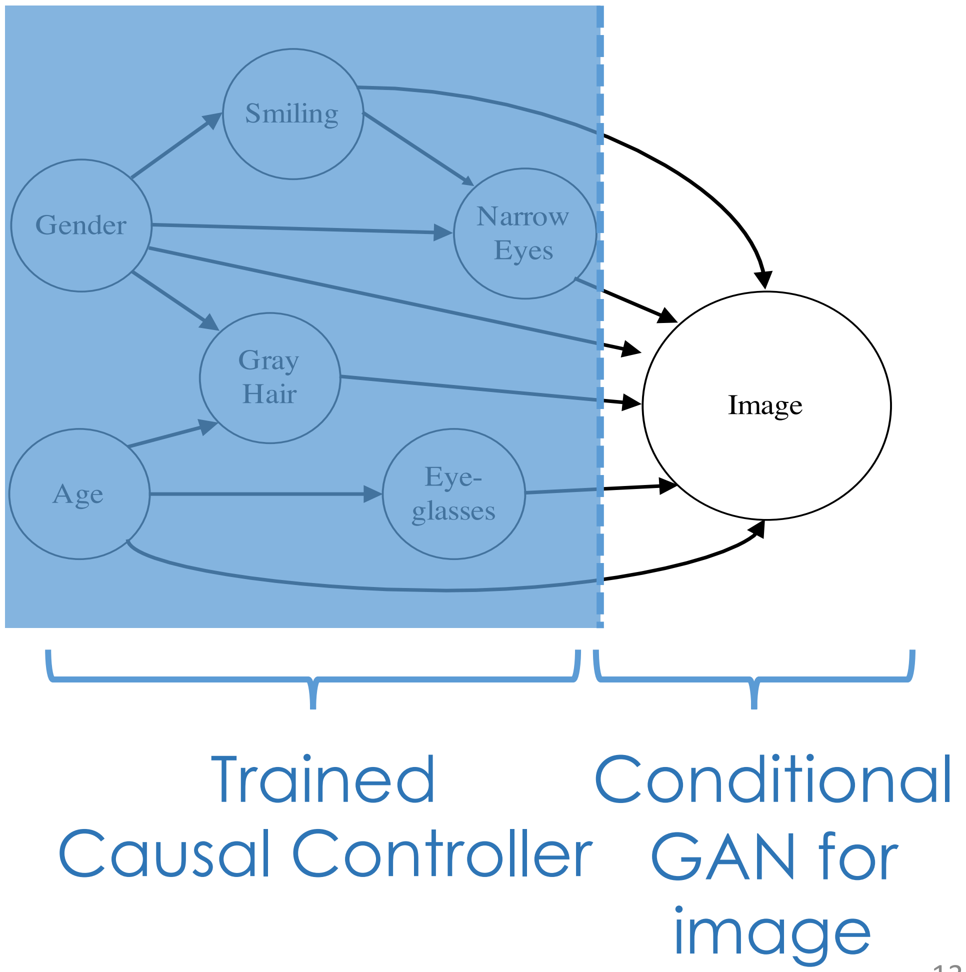


- **Theorem:**
  Correct causal graph + True observational distribution
  ➡ True Interventional distributions
  [under strict positivity]

CausalGAN: Causal Generative Model over Labels and Image

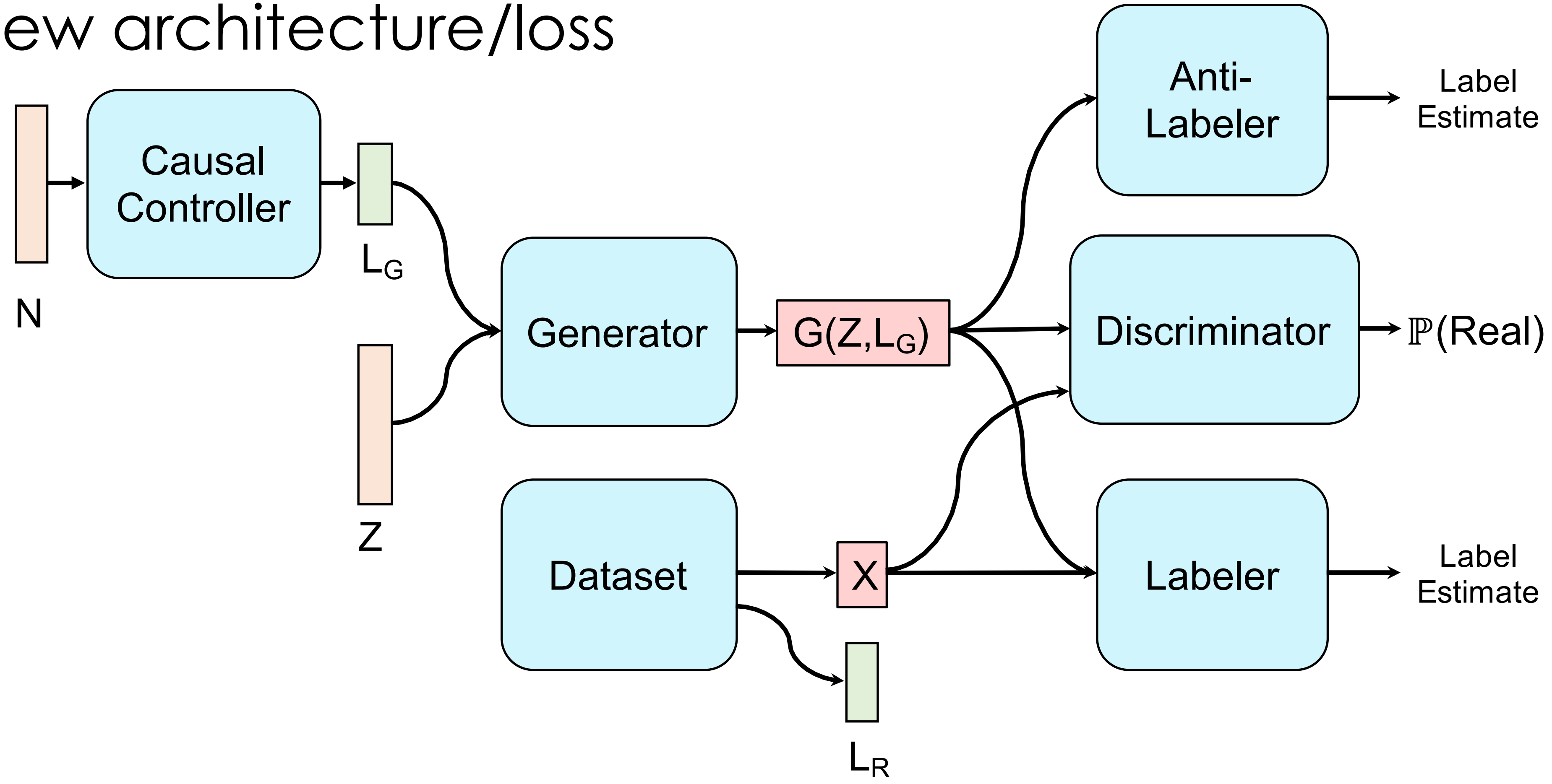
- Pre-train causal model over labels – *causal controller*
- Use a conditional GAN, given labels
- A new conditional GAN with theoretical guarantees

    [CGAN w/ same guarantee]



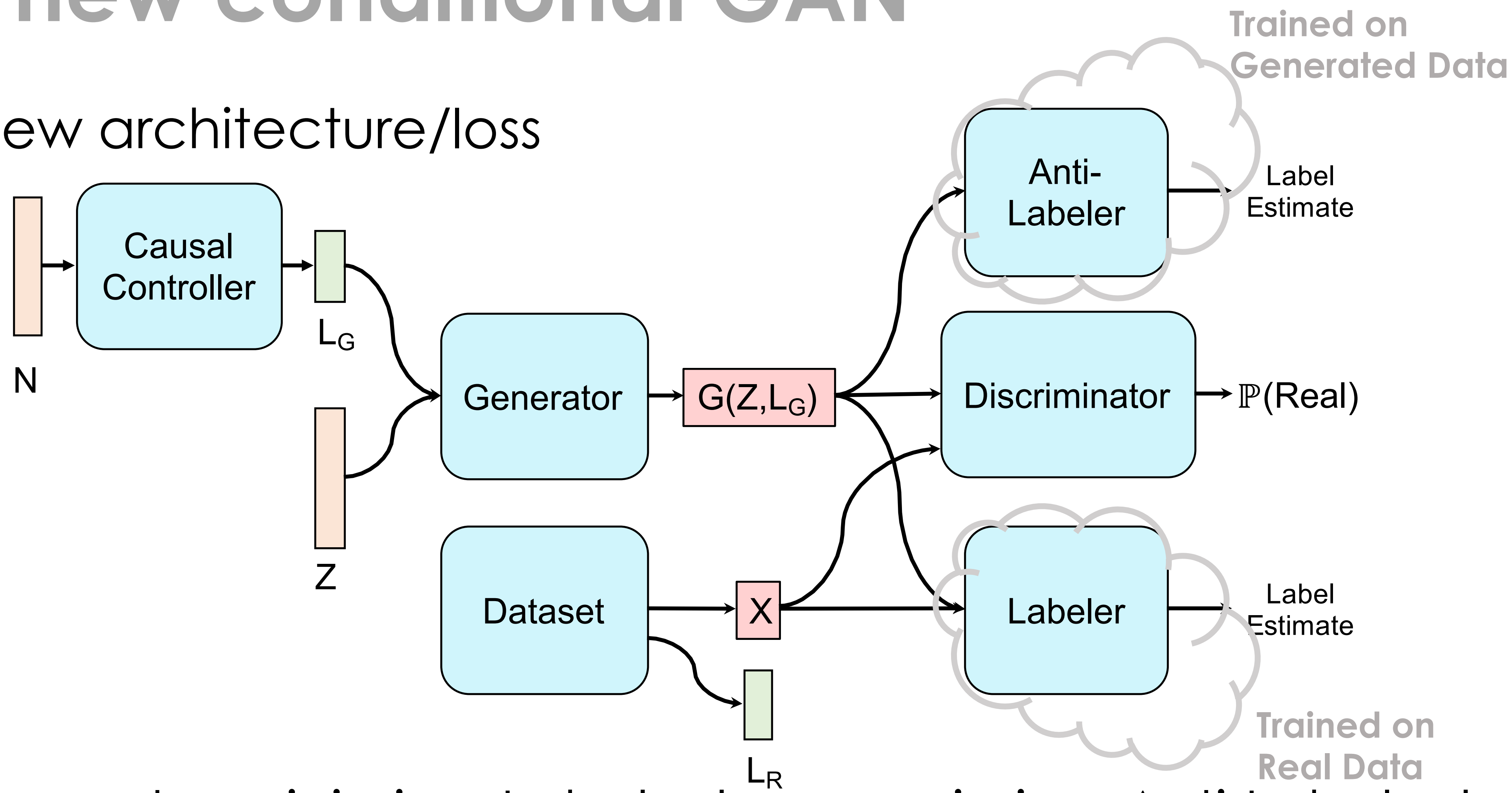
A new conditional GAN

- New architecture/loss



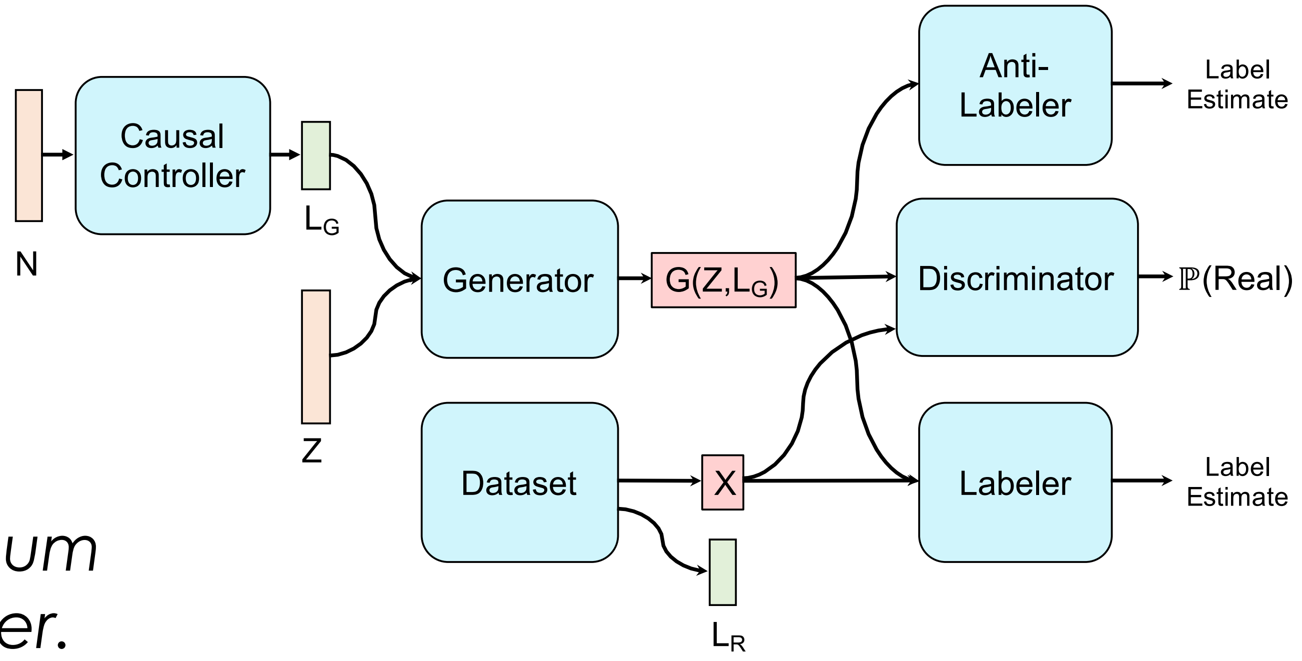
A new conditional GAN

- New architecture/loss



- Generator minimizes Labeler loss, maximizes Anti-Labeler loss

CausalGAN

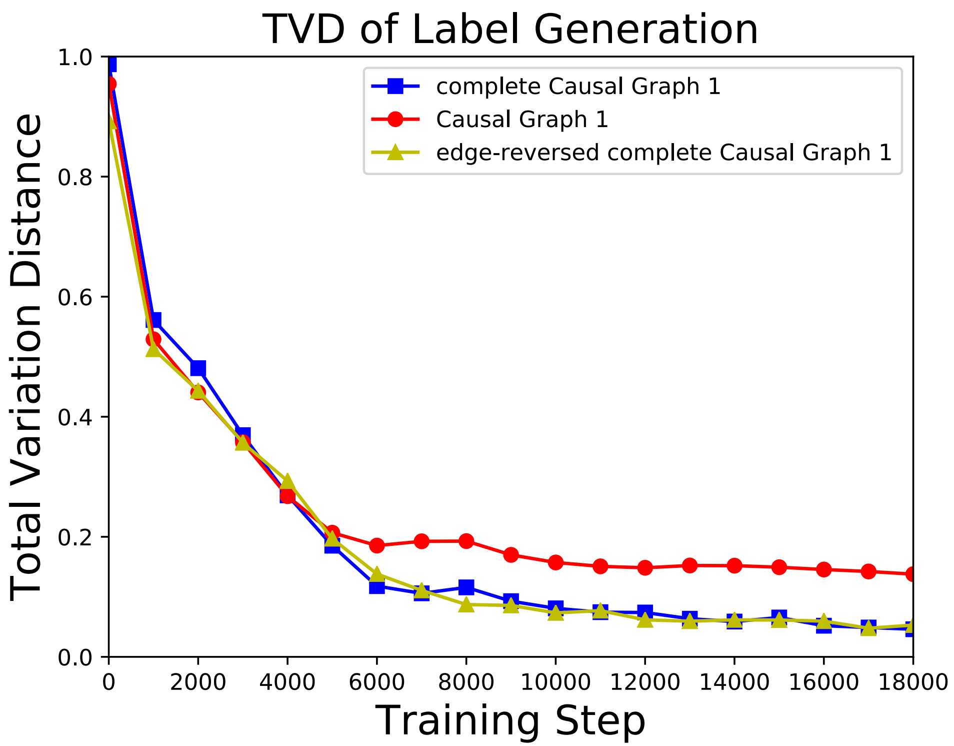


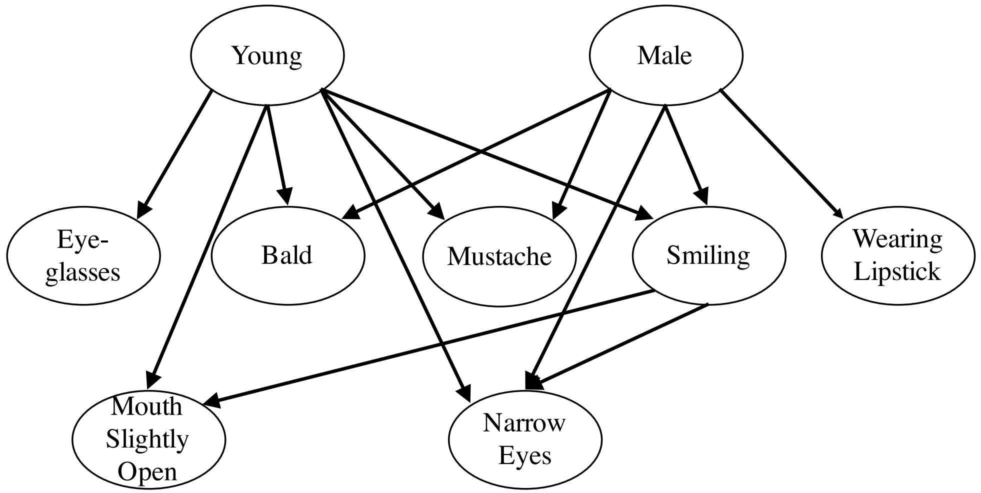
Theorem:
Optimize generator for the optimum Discriminator, Labeler, Anti-Labeler.

Then global optimal generator G^* samples from label conditioned image distributions:

$$\mathbb{P}(G^*(Z, l_G) = x) = \mathbb{P}(X = x | L_G = l_G)$$

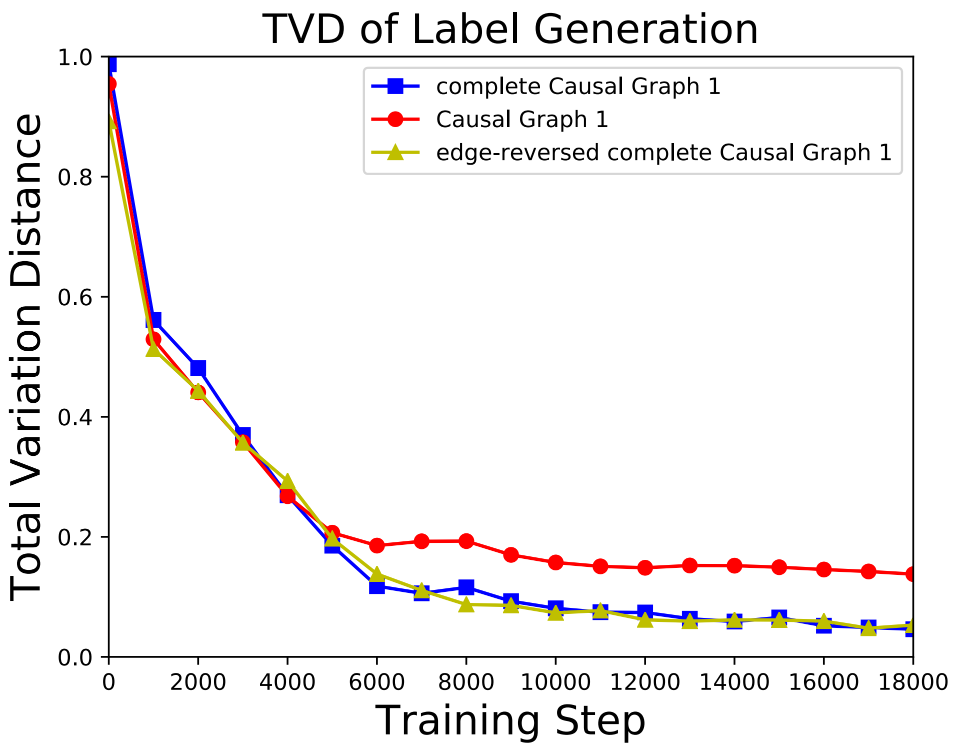
Results: Wasserstein GAN Training of Labels

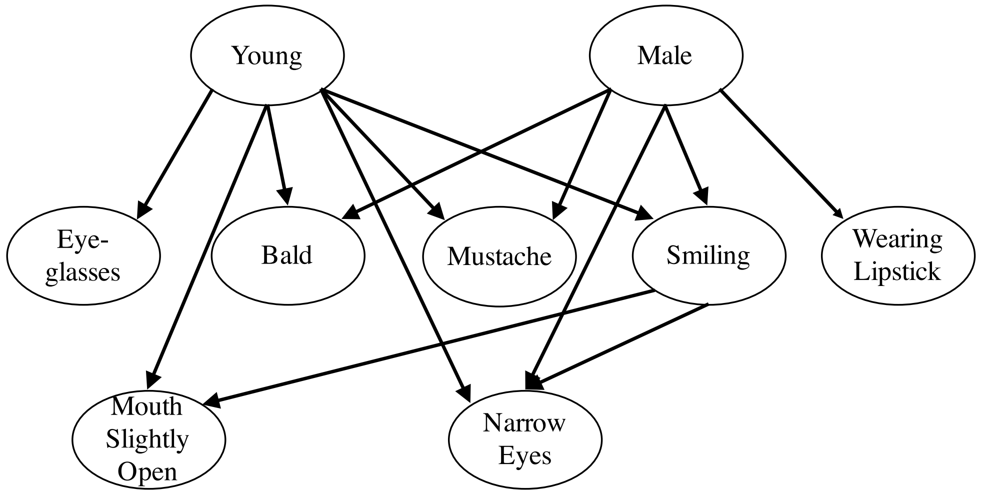




Causal Graph 1

Results: Wasserstein GAN Training of Labels

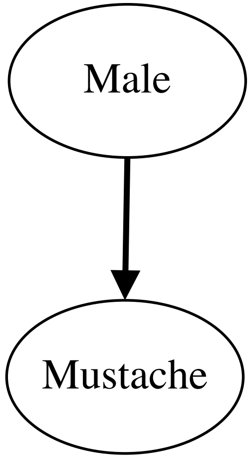




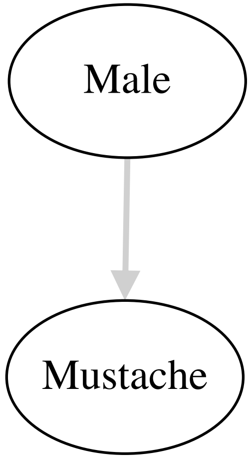
Causal Graph 1

Remark: Correctness of causal direction does not affect how well NNs can fit.

Results: CausalGAN



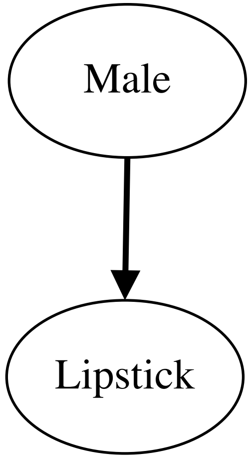
Conditioning on Mustache = 1



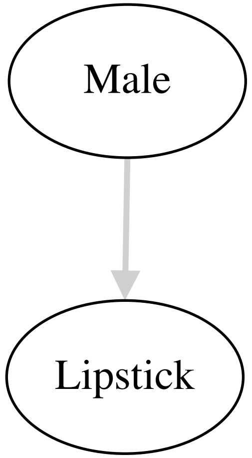
Intervening on Mustache = 1

$P(\text{Male} = 0 \mid do(\text{Mustache} = 1)) = P(\text{Male} = 0) \sim 0.6$

Results: CausalGAN



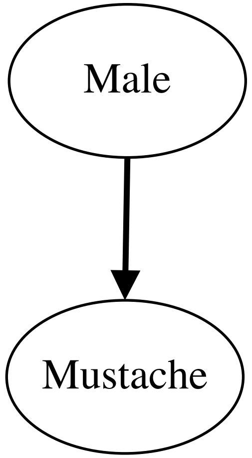
Conditioning on Lipstick = 1



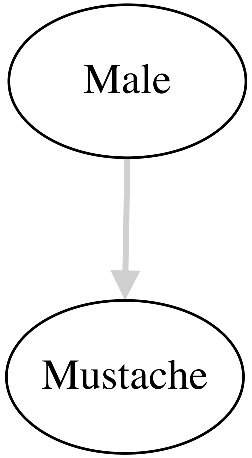
Intervening on Lipstick = 1

P(Male = 1 | do(Lipstick = 1)) = P(Male = 1) ~ 0.5

Results: CausalBEGAN

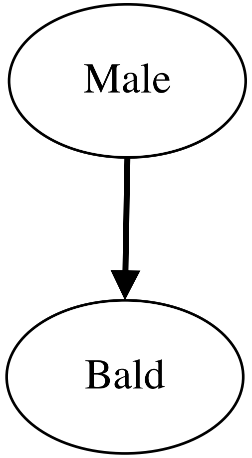


Conditioning on Mustache = 1

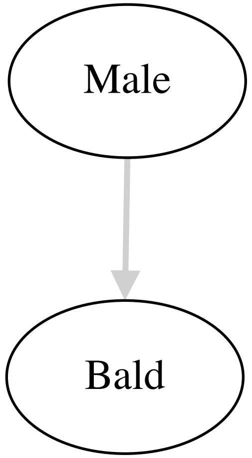


Intervening on Mustache = 1

Results: CausalBEGAN



Conditioning on Bald = 1



Intervening on Bald = 1

Questions?