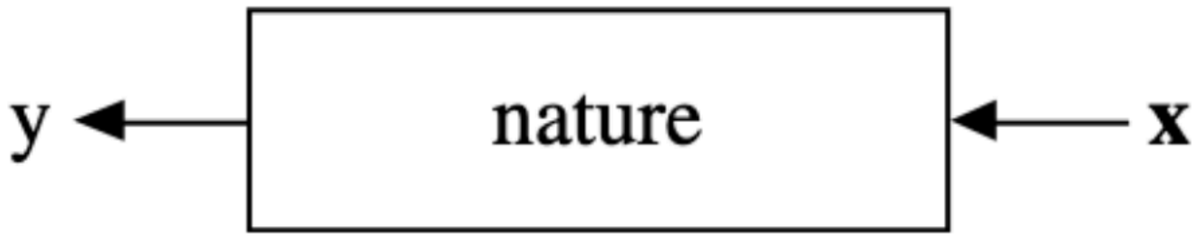
# Statistical Modeling: The Two Cultures

Sophia Lu

# Two Fundamentals of Statistical Order

- Data is generated by a black box:
  - $X$ (predictor) goes into the box
  - Generating process outputs $y$



# Two Fundamentals of Statistics

- **Prediction**: Predict what the responses are going to be for future input variables. **(Algorithmic modeling culture)**

# Two Fundamentals of Statistics

- **Prediction**: Predict what the responses are going to be for future input variables. **(Algorithmic modeling culture)**
- **Information**: To extract some information about how nature is associating the response variables to the input variables. **(Data modeling culture)**

# Data Modeling Culture

- Assumes a stochastic data generating process:
    - Example: $y_i \overset{iid}{\sim} f(X_i, \epsilon, \theta)$ where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $\epsilon$ is random noise, and $\theta$ are parameters.

# Data Modeling Culture

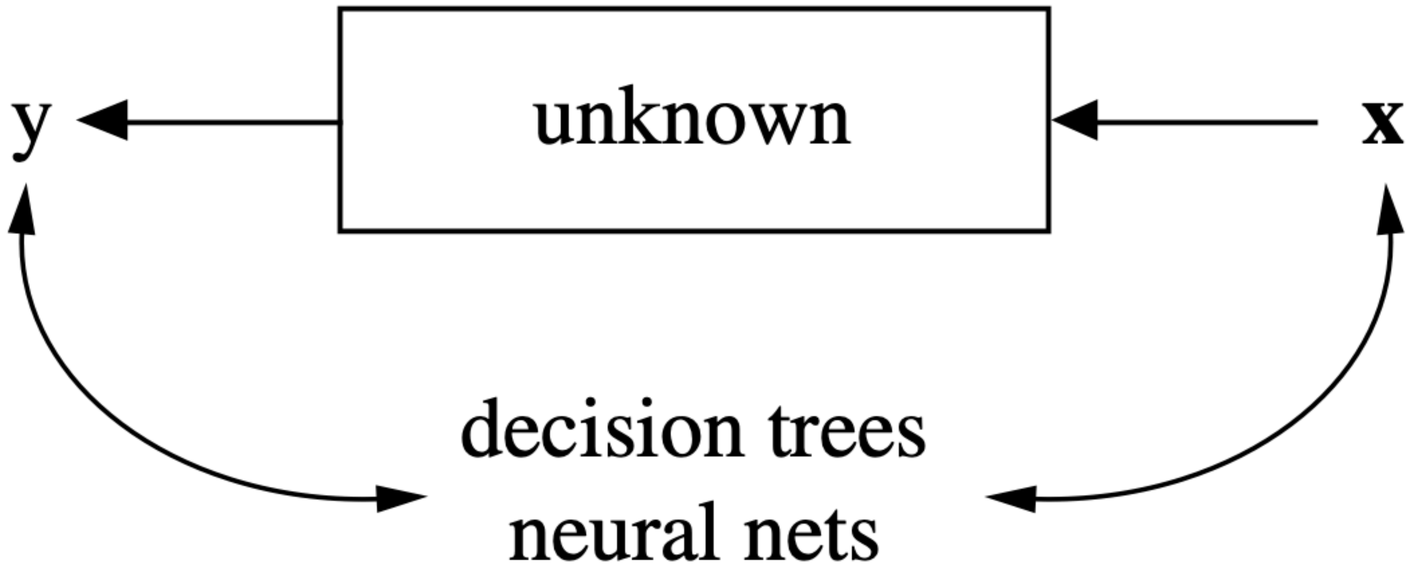
- Assumes a stochastic data generating process:
  - Example: $y_i \overset{iid}{\sim} f(X_i, \epsilon, \theta)$ where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $\epsilon$ is random noise, and $\theta$ are parameters.
  - Goal: Specify a model $M$ and estimate $\theta$ from $(X_i, y_i)_{i=1}^n$, $M$.

# Data Modeling Culture

- Assumes a stochastic data generating process:
  - Example: $y_i \overset{iid}{\sim} f(X_i, \epsilon, \theta)$ where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $\epsilon$ is random noise, and $\theta$ are parameters.
  - Goal: Specify a model $M$ and estimate $\theta$ from $(X_i, y_i)_{i=1}^n$, $M$.
  - Validation: Validate whether model is valid using hypothesis testing and residual analysis.

# Algorithmic Modeling Culture

- Considers inside of black box unknown.
- Goal: find $f(x)$ that predicts well without making distributional assumptions.



# Algorithmic Modeling Culture

- Considers inside of black box unknown.
- Goal: find $f(x)$ that predicts well without making distributional assumptions.
- Validation: select model via predictive accuracy / quantifiable measure of performance.

# In the 1990s…

- Breiman believed that 98% of statisticians subside within the data modeling culture, and a measly 2% of statisticians belong to the algorithmic modeling culture.

# In the 1990s…

- Breiman believed that 98% of statisticians subside within the data modeling culture, and a measly 2% of statisticians belong to the algorithmic modeling culture.
    - “Led to **irrelevant theory** and **questionable scientific conclusions**”.
    - “Kept statisticians from using more suitable algorithmic models”.
    - **“**Prevented statisticians from working on exciting new problems”.

# In the 1990s…

- “[…] I started reading the *Annals of Statistics*, the flagship journal of theoretical statistics, and was **bemused**. Every article started with

# In the 1990s…

- “[…] I started reading the *Annals of Statistics*, the flagship journal of theoretical statistics, and was **bemused**. Every article started with
- > Assume that the data are generated by the following model: …

# In the 1990s…

- "[…] I started reading the *Annals of Statistics*, the flagship journal of theoretical statistics, and was **bemused**. Every article started with

  > Assume that the data are generated by the following model: …

- followed by mathematics exploring inference, hypothesis testing and asymptotics. There is a wide spectrum of opinion regarding the usefulness of the theory published in the *Annals of Statistics* to the field of statistics as a science that **deals with data**. […]

# In the 1990s…

- “[…] I started reading the *Annals of Statistics*, the flagship journal of theoretical statistics, and was **bemused**. Every article started with

> Assume that the data are generated by the following model: . . .

- followed by mathematics exploring inference, hypothesis testing and asymptotics. There is a wide spectrum of opinion regarding the usefulness of the theory published in the *Annals of Statistics* to the field of statistics as a science that **deals with data**. […]

- Still, there have been some gems that have combined nice theory and significant applications. ”

# In the 1990s…

- “[…] Even in applications, data models are universal. For instance, in the *Journal of the American Statistical Association (JASA)*, virtually every article contains a statement of the form:

# In the 1990s…

- “[…] Even in applications, data models are universal. For instance, in the *Journal of the American Statistical Association (JASA)*, virtually every article contains a statement of the form:

> Assume that the data are generated by the following model: . . .

# In the 1990s…

- “[…] Even in applications, data models are universal. For instance, in the *Journal of the American Statistical Association (JASA)*, virtually every article contains a statement of the form:

> Assume that the data are generated by the follow-
> ing model: . . .

- I am deeply troubled by the current and past use of data models in applications, where quantitative conclusions are drawn and perhaps policy decisions made.”

# Breiman’s Criticism

- When data is gathered from complex systems, modeling the data generating process by a **simple** parametric model results in questionable conclusions.

# Breiman’s Three Lessons

- Rashomon: multiplicity of good models.
- Occam: conflict between simplicity and accuracy.
- Bellman: dimensionality – curse or blessing.

# Breiman's Three Lessons

- Rashomon: multiplicity of good models.
- Occam: conflict between simplicity and accuracy.
- Bellman: dimensionality – curse or blessing.

# LassoNet: A Neural Network with Feature Sparsity

# LassoNet: A Neural Network with Feature Sparsity

- Traditional neural networks often lack interpretability, especially when dealing with high-dimensional data.

# LassoNet: A Neural Network with Feature Sparsity

- Traditional neural networks often lack interpretability, especially when dealing with high-dimensional data.
- One might ask:
    - Are there redundant features?
    - What are the most effective / representative features to characterize a specific disease?

# LassoNet: A Neural Network with Feature Sparsity

- Traditional neural networks often lack interpretability, especially when dealing with high-dimensional data.
- Feature selection is crucial:
    - Improve interpretability
    - Reduce overfitting
    - Enhance computational efficiency

# LassoNet: A Neural Network with Feature Sparsity

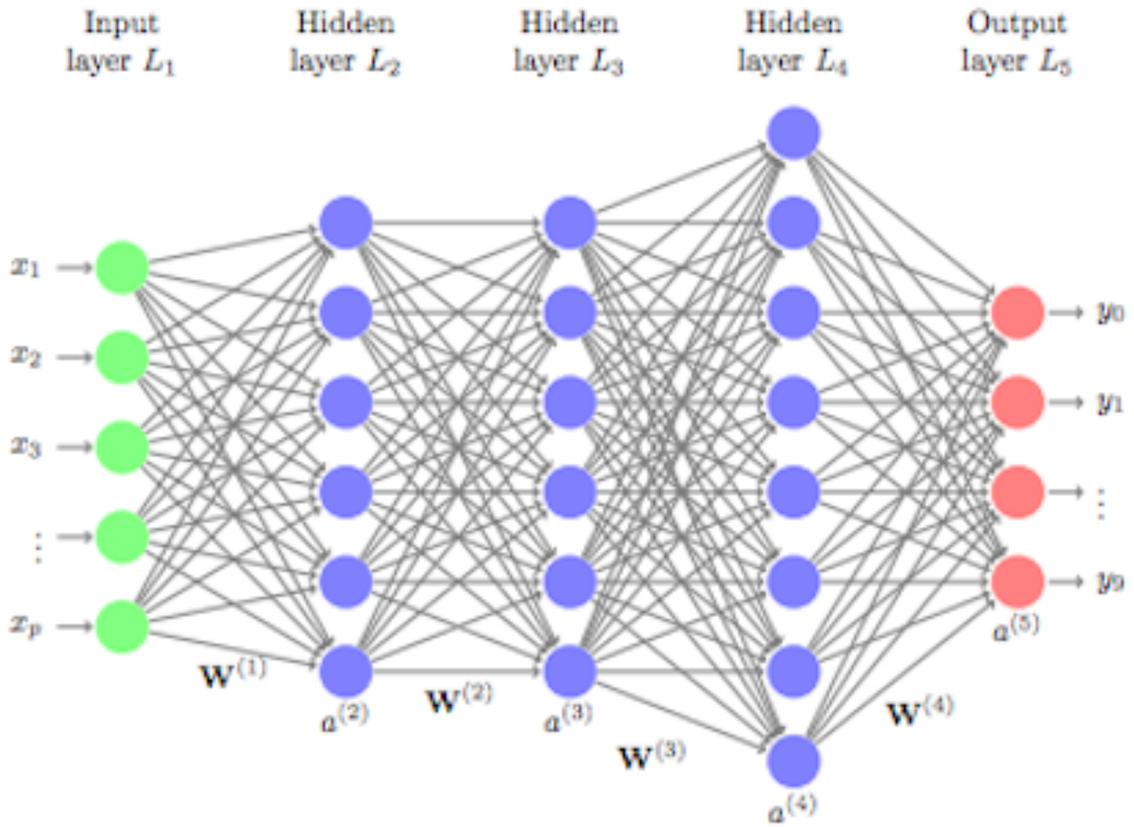
- Traditional neural networks often lack interpretability, especially when dealing with high-dimensional data.
- Feature selection is crucial:
    - Improve interpretability
    - Reduce overfitting
    - Enhance computational efficiency
- **Challenge**: Unlike linear models (LASSO), neural networks do not inherently support global feature selection.

# LassoNet: A Neural Network with Feature Sparsity

- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.

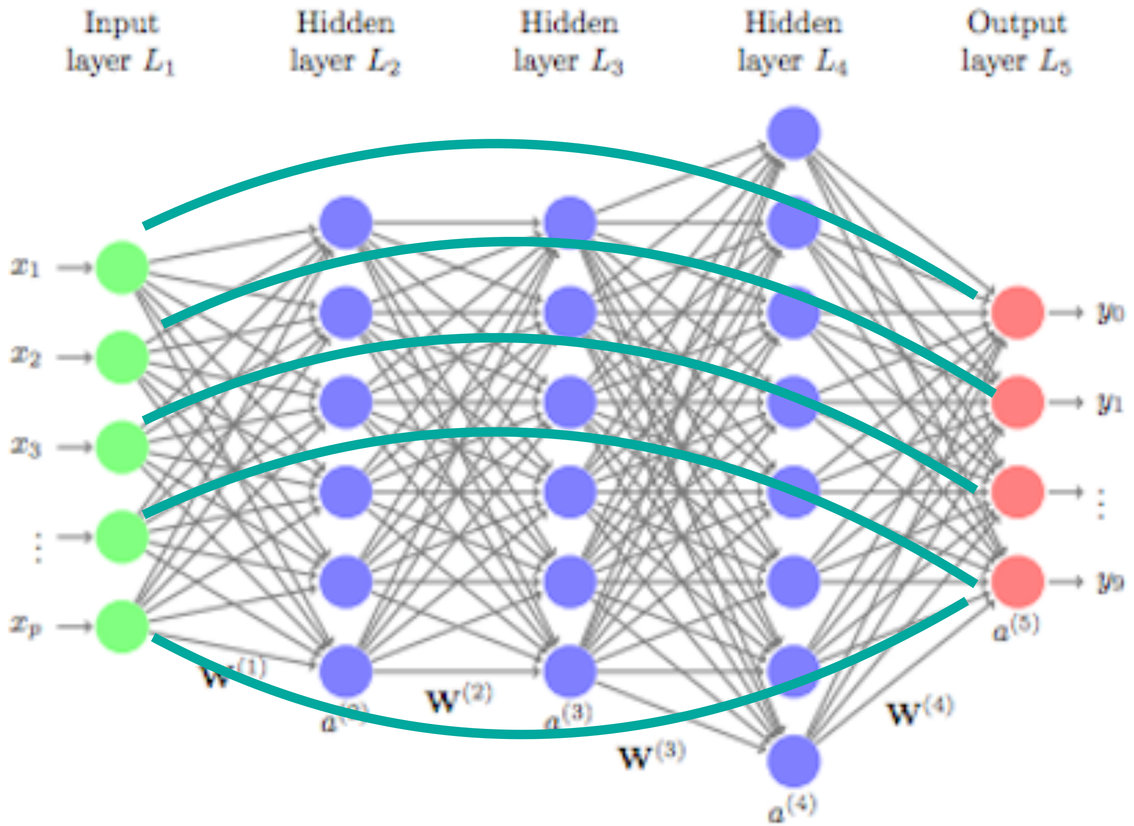
# LassoNet: A Neural Network with Feature Sparsity

- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.



# LassoNet: A Neural Network with Feature Sparsity

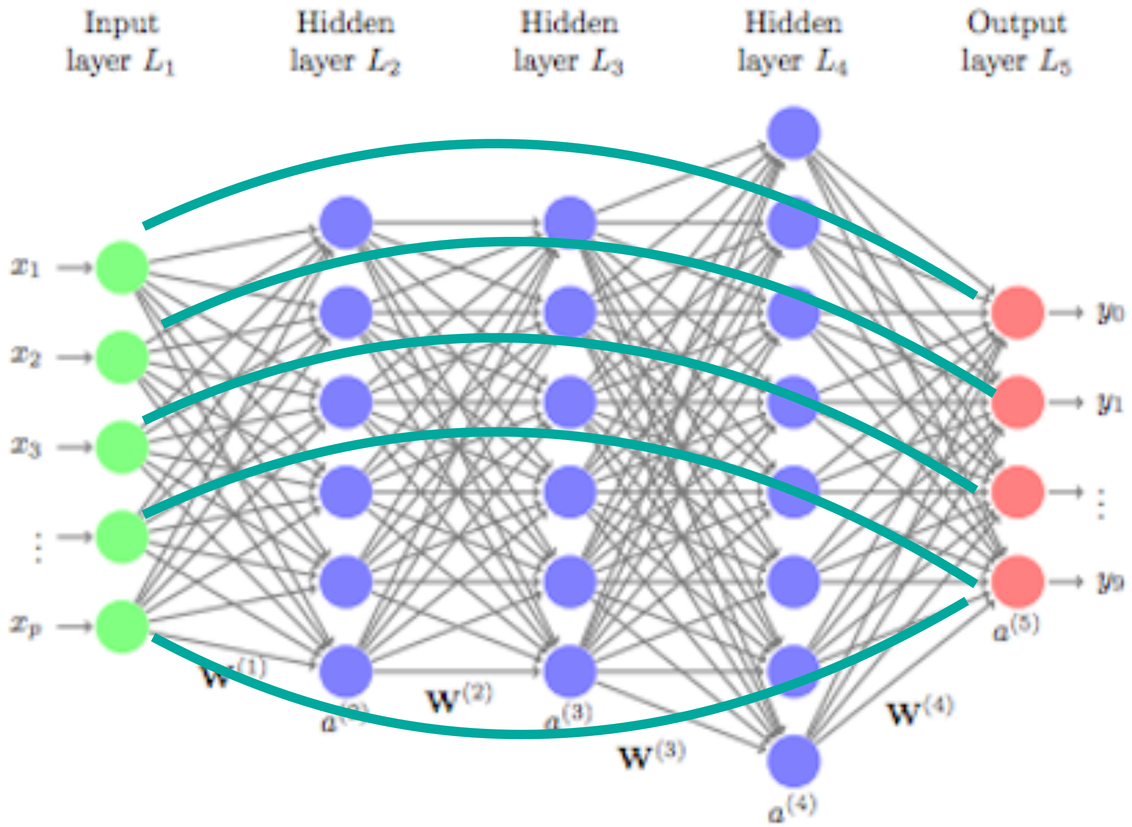
- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.




**skip connections**

# LassoNet: A Neural Network with Feature Sparsity

- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.




**Key constraint:**
A feature can only influence hidden layers if its corresponding skip connection (with linear coefficients) is active

# LassoNet: A Neural Network with Feature Sparsity

- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.
- Formally, objective function is as follows:

$$\min_{\theta, W} \quad L(\theta, W) + \lambda \|\theta\|_1$$

$$\text{s.t.} \ \|W_j^{(1)}\|_\infty \leq M|\theta_j|, j = 1, \ldots, d$$

# LassoNet: A Neural Network with Feature Sparsity

- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.
- Formally, objective function is as follows:

$$\min_{\theta, W} \quad L(\theta, W) + \lambda \|\theta\|_1 \qquad \text{Lasso penalty}$$

$$\text{s.t.} \ \|W_j^{(1)}\|_\infty \leq M|\theta_j|, j = 1, \ldots, d$$

# LassoNet: A Neural Network with Feature Sparsity

- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.
- Formally, objective function is as follows:

$$\min_{\theta, W} \quad L(\theta, W) + \lambda \|\theta\|_1$$

$$\text{s.t.} \ \|W_j^{(1)}\|_\infty \leq M|\theta_j|, j = 1, \ldots, d$$

Controls total amount of non-linearity involving feature $j$ according to relative effect importance of $X_j$.

# LassoNet: A Neural Network with Feature Sparsity

- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.
- Formally, objective function is as follows:

$$\min_{\theta, W} \quad L(\theta, W) + \lambda \|\theta\|_1$$

$$\text{s.t.} \ \|W_j^{(1)}\|_\infty \leq M|\theta_j|, j = 1, \ldots, d$$

If $M = 0$, all hidden units are inactive and only skip connection remains.

**Exact Lasso.**

# LassoNet: A Neural Network with Feature Sparsity

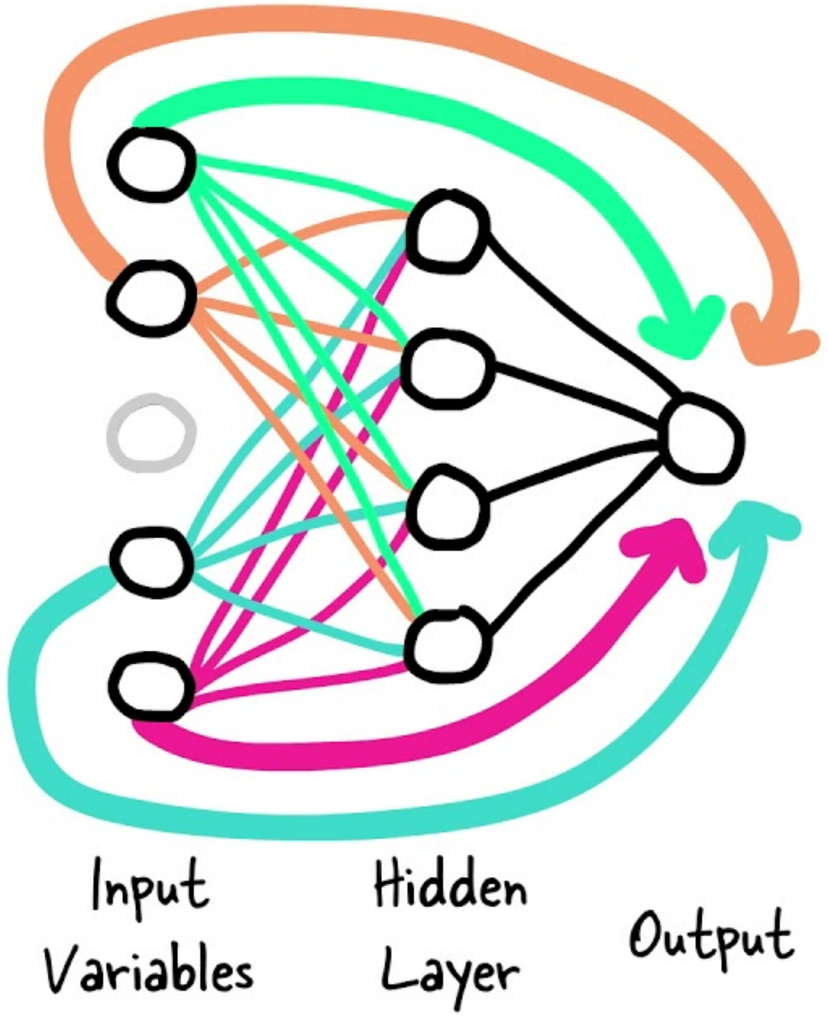
- **LassoNet** extends Lasso regression and feature sparsity to feedforward neural networks.
- Formally, objective function is as follows:

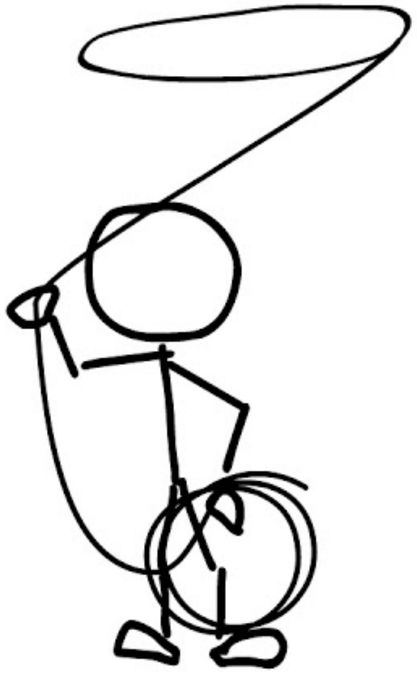
$$\min_{\theta, W} \quad L(\theta, W) + \lambda \|\theta\|_1$$

$$\text{s.t.} \ \|W_j^{(1)}\|_\infty \leq M|\theta_j|, j = 1, \ldots, d$$

If $M = \infty$, recovers standard unregularized feed-forward neural network.

# LassoNet: A Neural Network with Feature Sparsity



# LassoNet: A Neural Network with Feature Sparsity

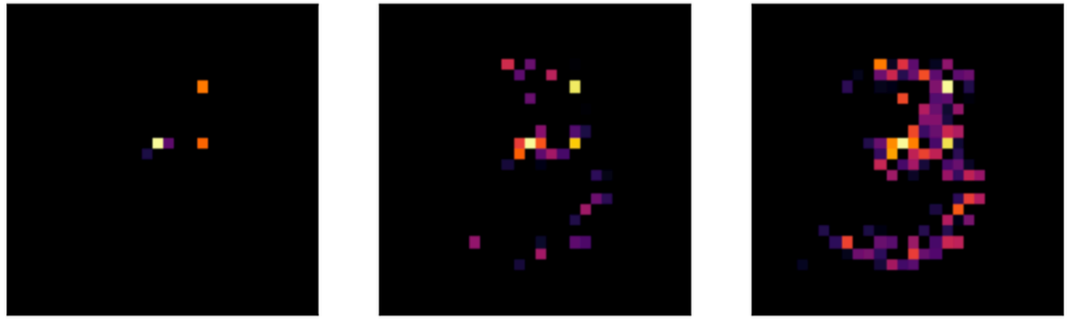


*Figure 7.* Results for LassoNet in choosing the most informative pixels of images of the digit 3 in the MNIST dataset, for three different penalty levels ($\lambda = 5, \lambda = 1, \lambda = 0.1$).

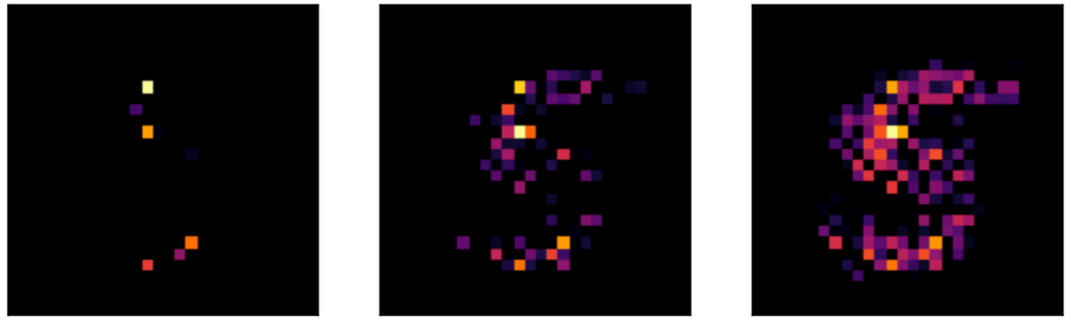


*Figure 8.* Results for LassoNet in choosing the most informative pixels of images of the digit 5 in the MNIST dataset, for the three penalty levels.

# Breiman’s Three Lessons

- Rashomon: multiplicity of good models.
- Occam: conflict between simplicity and accuracy.
- Bellman: dimensionality – curse or blessing.

# Breiman's Three Lessons

- When a model is fit to data to draw quantitate conclusions:
  - "The conclusions are about the model's mechanism, and not about nature's mechanism."
- Thus,
  - "If the model is a poor emulation of nature, the conclusions may be wrong."

# Efficient Generative Modeling via Penalized Optimal Transport Network

# Synthetic Data Generation

- **Why are synthetic data important?**

# Synthetic Data Generation

- Generation of **high-quality** synthetic data is important across many disciplines.
- Researchers and practitioners may want to use synthetic data for
  - Model evaluation
  - Data augmentation
  - Model selection

# Synthetic Data Generation

- Generation of **high-quality** synthetic data is important across many disciplines.
- Researchers and practitioners may use synthetic data for
  - **Model evaluation:** When the true parameter of interest is unobserved (e.g., causal effects), researchers typically rely on Monte Carlo studies with discretionary assumptions to evaluate proposed methods. 
  - Data augmentation
  - Model selection

Athey, Susan, et al. "Using wasserstein generative adversarial networks for the design of monte carlo simulations." *Journal of Econometrics*

# Synthetic Data Generation

- Generation of **high-quality** synthetic data is important across many disciplines.
- Researchers and practitioners may use synthetic data for
  - **Model evaluation:** Faithful synthetic data allows for systematic evaluation of methods under conditions that closely reflect real-world scenarios.
  - Data augmentation
  - Model selection

Athey, Susan, et al. "Using wasserstein generative adversarial networks for the design of monte carlo simulations." *Journal of Econometrics*

# Synthetic Data Generation

- Generation of **high-quality** synthetic data is important across many disciplines.
- Researchers and practitioners may use synthetic data for
  - Model evaluation
  - **Data augmentation:** In likelihood-free inference, use synthetic data to directly approximate posterior and bypass likelihood evaluation. In machine learning, address class imbalance issues by adding synthetic data of minor groups.
  - Model selection

Frid-Adar, Maayan, et al. "Synthetic data augmentation using GAN for improved liver lesion classification." *2018 IEEE 15th international symposium on biomedical imaging.*

# Synthetic Data Generation

- Generation of **high-quality** synthetic data is important across many disciplines.
- Researchers and practitioners may use synthetic data for
    - Model evaluation
    - Data augmentation
    - **Model selection:** Evaluate empirical performance of estimators on synthetic datasets, and use these performance metrics to guide model selection process.

Ganganwar, Vaishali, and Ratnavel Rajalakshmi. "Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification." *Journal of Information and Telecommunication*

# Synthetic Data Generation

- Generation of **high-quality** synthetic data is important across many disciplines.
- Researchers and practitioners may use synthetic data for
    - Model evaluation
    - Data augmentation
    - Model selection
- Goal: generating **high-quality** synthetic data whose distributions closely mirror the true data-generating mechanism.

# Generative Modeling

- **Generative models** aim to approximate the underlying probability distribution of observed data and produce new samples from the approximated distribution.

- Wasserstein Generative Adversarial Networks (WGANs) are a popular and powerful tool among generative models for synthetic data generation.

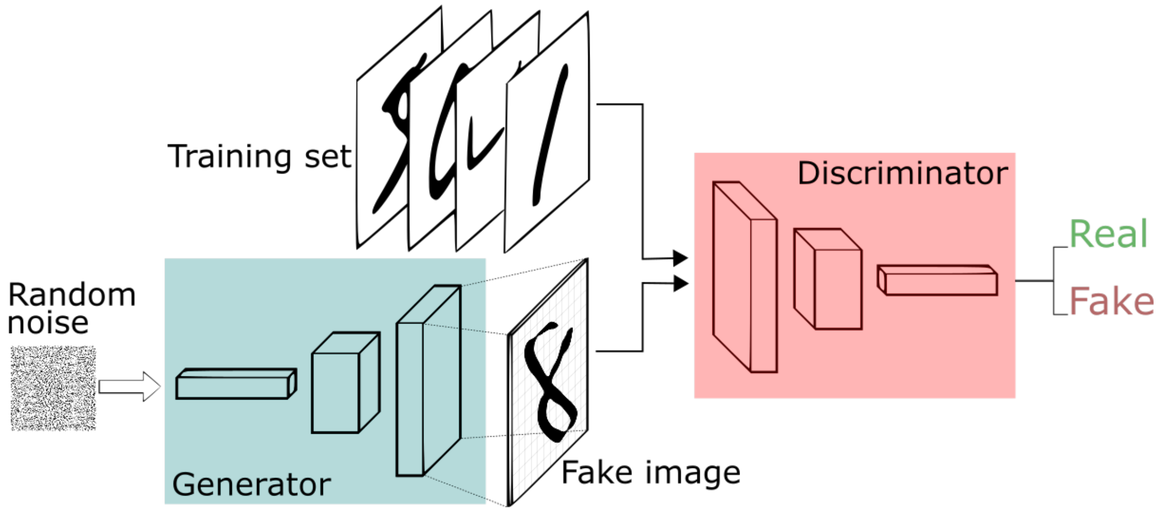
# Generative Modeling

- Generative models aim to approximate the underlying probability distribution of observed data and produce new samples from the approximated distribution.
- **Wasserstein Generative Adversarial Networks (WGANs)** are a popular and powerful tool among generative models for synthetic data generation.

# Generative Modeling

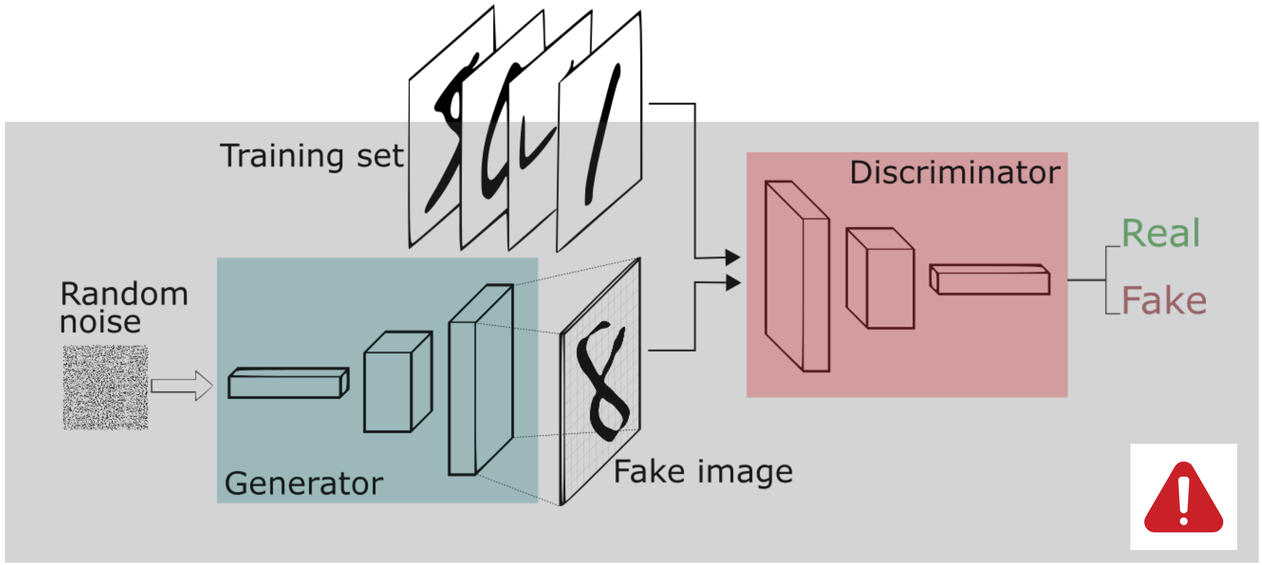
- Generative models aim to approximate the underlying probability distribution of observed data and produce new samples from the approximated distribution.

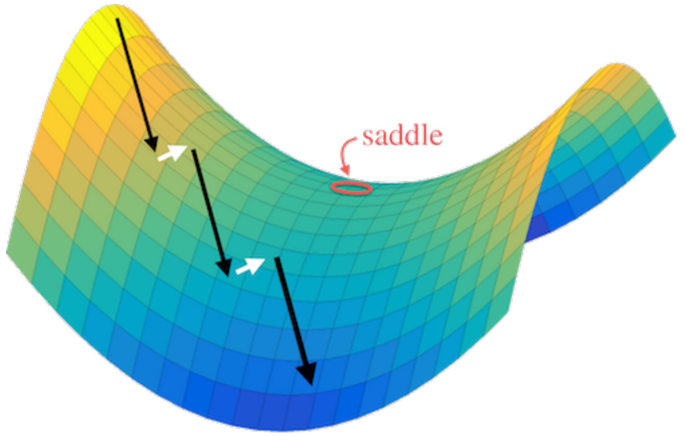
- **Wasserstein Generative Adversarial Networks (WGANs)** are a popular and powerful tool among generative models for synthetic data generation.



# Generative Modeling

- Generative models aim to approximate the underlying probability distribution of observed data and produce new samples from the approximated distribution.
- **Wasserstein Generative Adversarial Networks (WGANs)** are a popular and powerful tool among generative models for synthetic data generation.



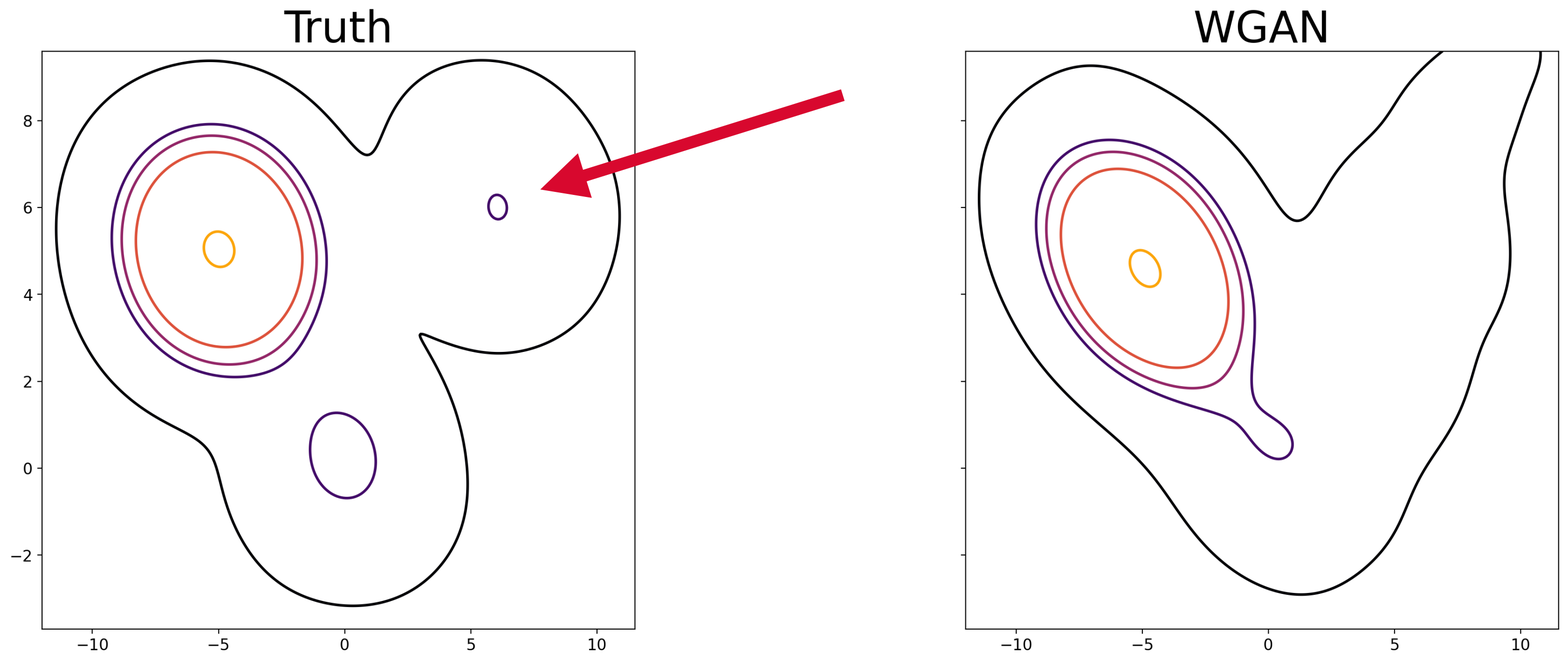



# The Mode Collapse Pathology

- **Mode collapse:** the generator only produces a limited subset of outputs (reduced diversity).

# The Mode Collapse Pathology

- **Mode collapse:** the generator only produces a limited subset of outputs (reduced diversity).
  - **Type I mode collapse (mode dropping):** failing to assign adequate weights to clusters.

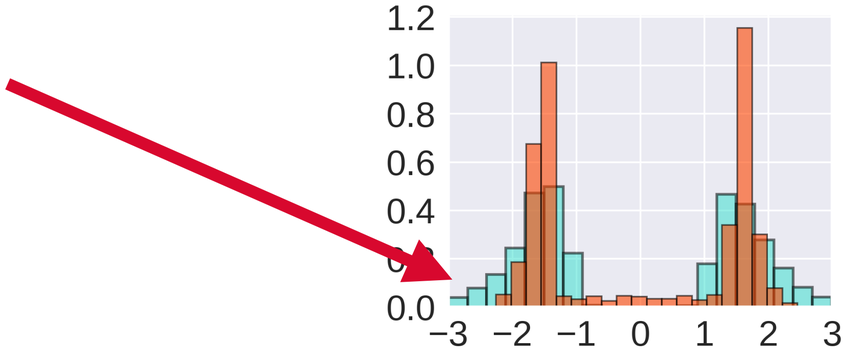


# The Mode Collapse Pathology

- **Mode collapse:** the generator only produces a limited subset of outputs (reduced diversity).
    - **Type I mode collapse (mode dropping):** failing to assign adequate weights to clusters.
        - Practical implications: significant misrepresentation of subpopulations in heterogeneous populations.

# The Mode Collapse Pathology

- **Mode collapse:** the generator only produces a limited subset of outputs (reduced diversity).
  - **Type I mode collapse (mode dropping):** failing to assign adequate weights to clusters.
    - Practical implications: significant misrepresentation of subpopulations in heterogeneous populations.
  - **Type II mode collapse (support shrinkage):** failing to adequately represent the tail behavior of the data-generating distribution.



# The Mode Collapse Pathology

- **Mode collapse:** the generator only produces a limited subset of outputs (reduced diversity).
    - **Type I mode collapse (mode dropping):** failing to assign adequate weights to clusters.
        - Practical implications: significant misrepresentation of subpopulations in heterogeneous populations.
    - **Type II mode collapse (support shrinkage):** failing to adequately represent the tail behavior of the data-generating distribution.
        - Practical implications: omission of important extremal information.

# The Mode Collapse Pathology

- **Causes of mode collapse:**
    - Instability of Wasserstein distance in high-dimensions.
    - Training instability from adversarial training process (minimax optimization).
- If we do not regularize in the right way, the model will be prone to mode collapse.

# The Mode Collapse Pathology

- **Causes of mode collapse:**
    - Instability of Wasserstein distance in high-dimensions.
    - Training instability from adversarial training process (minimax optimization).
- **If we do not regularize in the right way, the model will be prone to mode collapse.** 

# The POTNet Framework

## **P**enalized **O**ptimal **T**ransport **Net**work

- Since marginal distributions converge rapidly (no dependence on dimension), leverage marginal penalty as a form of regularization to guide alignment of joint distributions.
    - This formulation offers a more interpretable loss function.
- Use Monge-Kantorovich **primal formulation** as opposed to dual formulation, eliminating the need for discriminator network and adversarial training process.

# The POTNet Framework

## **P**enalized **O**ptimal **T**ransport **Net**work

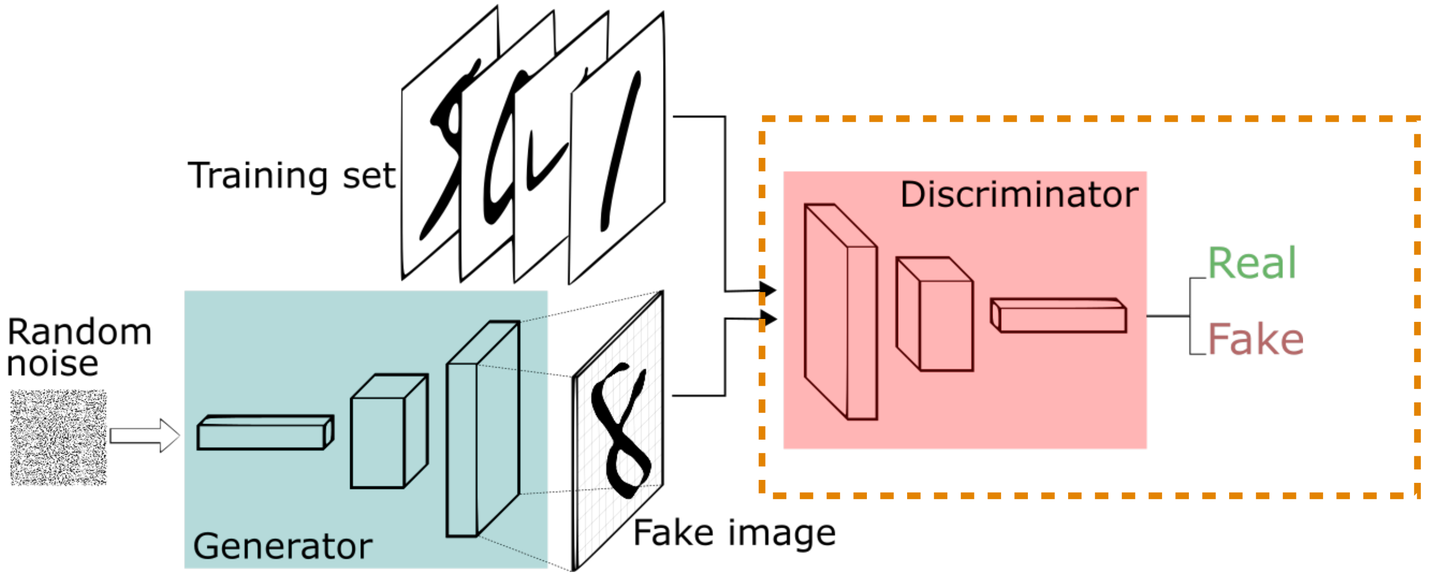
- Since marginal distributions converge rapidly (no dependence on dimension), leverage marginal penalty as a form of regularization to guide alignment of joint distributions.
    - This formulation offers a more interpretable loss function.
- Use Monge-Kantorovich **primal formulation** as opposed to dual formulation, eliminating the need for discriminator network and adversarial training process.

# The POTNet Framework

## **P**enalized **O**ptimal **T**ransport **Net**work

- Since marginal distributions converge rapidly (no dependence on dimension), leverage marginal penalty as a form of regularization to guide alignment of joint distributions.

- Use Monge-Kantorovich **primal formulation** as opposed to dual formulation, eliminating the need for discriminator network and adversarial training process.

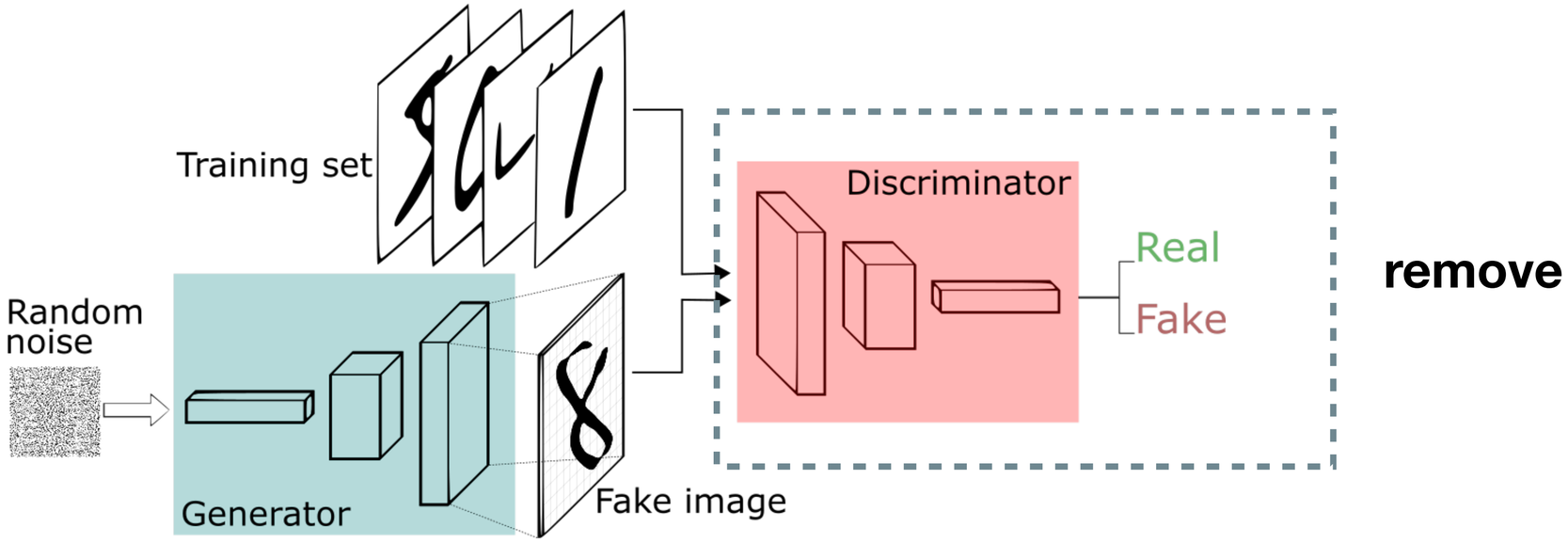


# The POTNet Framework

## **P**enalized **O**ptimal **T**ransport **Net**work

- Since marginal distributions converge rapidly (no dependence on dimension), leverage marginal penalty as a form of regularization to guide alignment of joint distributions.

- Use Monge-Kantorovich **primal formulation** as opposed to dual formulation, eliminating the need for discriminator network and adversarial training process.



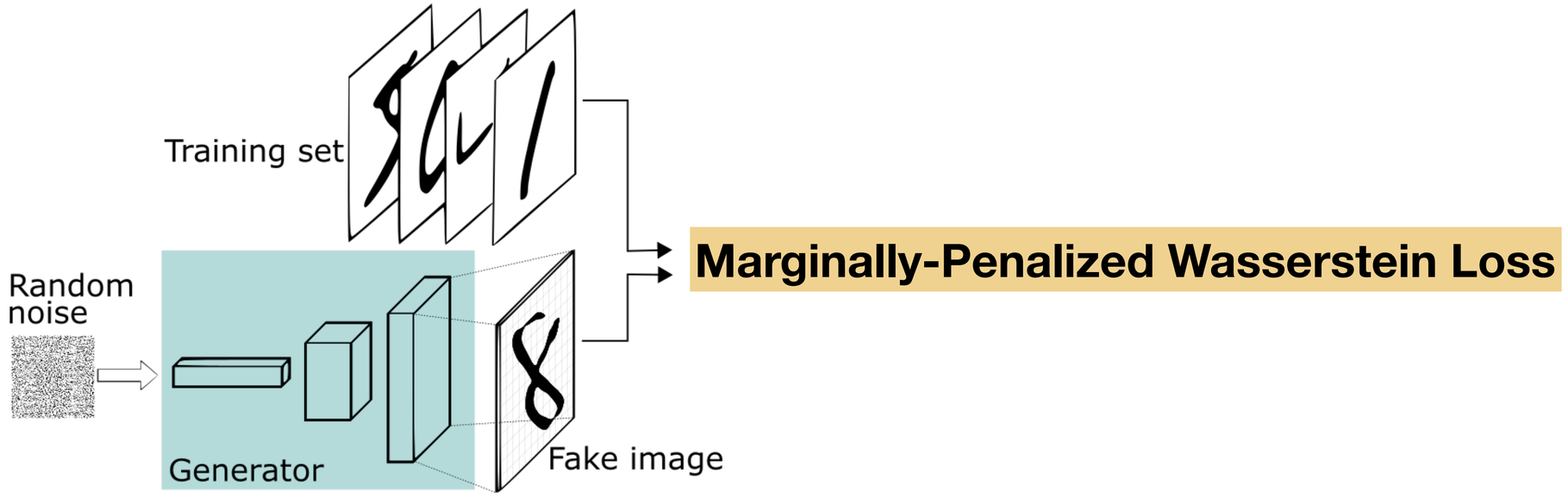

**remove**

# The POTNet Framework

## **P**enalized **O**ptimal **T**ransport **Net**work

- Since marginal distributions converge rapidly (no dependence on dimension), leverage marginal penalty as a form of regularization to guide alignment of joint distributions.

- Use Monge-Kantorovich **primal formulation** as opposed to dual formulation, eliminating the need for discriminator network and adversarial training process.

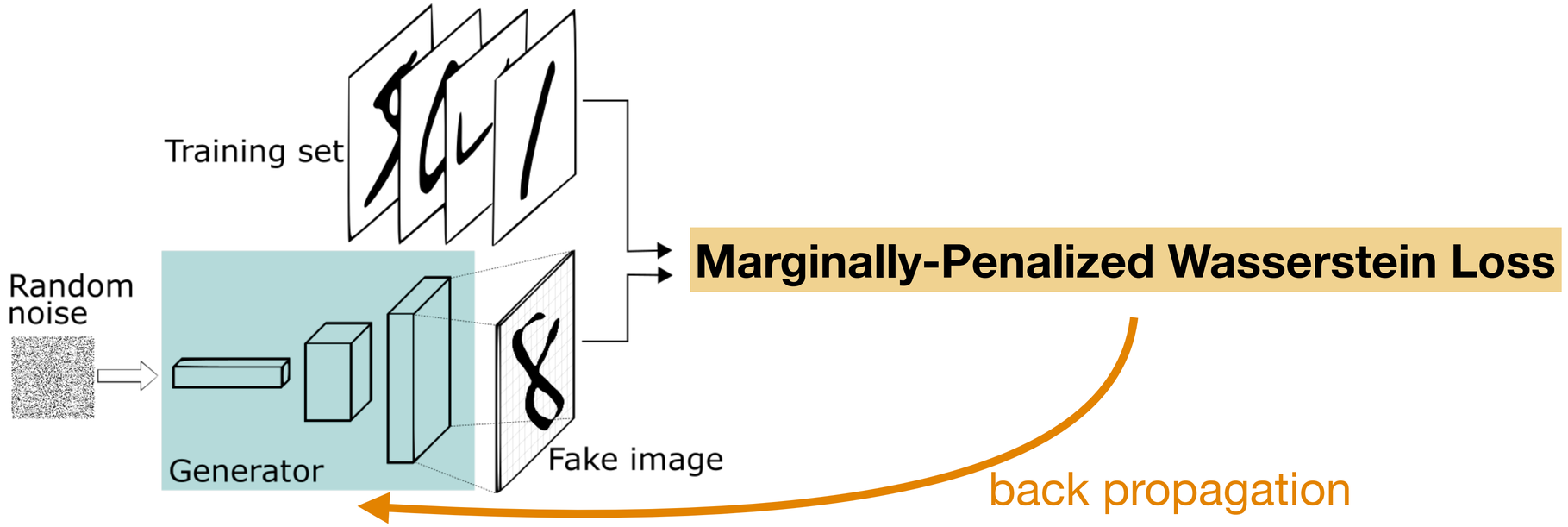


# The POTNet Framework

## **P**enalized **O**ptimal **T**ransport **Net**work

- Since marginal distributions converge rapidly (no dependence on dimension), leverage marginal penalty as a form of regularization to guide alignment of joint distributions.

- Use Monge-Kantorovich **primal formulation** as opposed to dual formulation, eliminating the need for discriminator network and adversarial training process.



# The POTNet Framework

## Marginally-Penalized Wasserstein Distance

**Marginally-Penalized Wasserstein (MPW) Distance**

$$\mathcal{D}(\mu,\nu) := \underbrace{W_1(\mu,\nu)}_{joint\ distance} + \sum_{j=1}^{d} \underbrace{\lambda_j W_1((p_j)_*\mu, (p_j)_*\nu)}_{marginal\ penalty}$$

# The POTNet Framework

## Marginally-Penalized Wasserstein Distance

**Marginally-Penalized Wasserstein (MPW) Distance**

$$\mathcal{D}(\mu,\nu) := \underbrace{W_1(\mu,\nu)}_{joint\ distance} + \sum_{j=1}^{d} \underbrace{\lambda_j W_1((p_j)_*\mu, (p_j)_*\nu)}_{marginal\ penalty}$$

regularization factor on the $j$th marginal

# The POTNet Framework

## Marginally-Penalized Wasserstein Distance

**Marginally-Penalized Wasserstein (MPW) Distance**

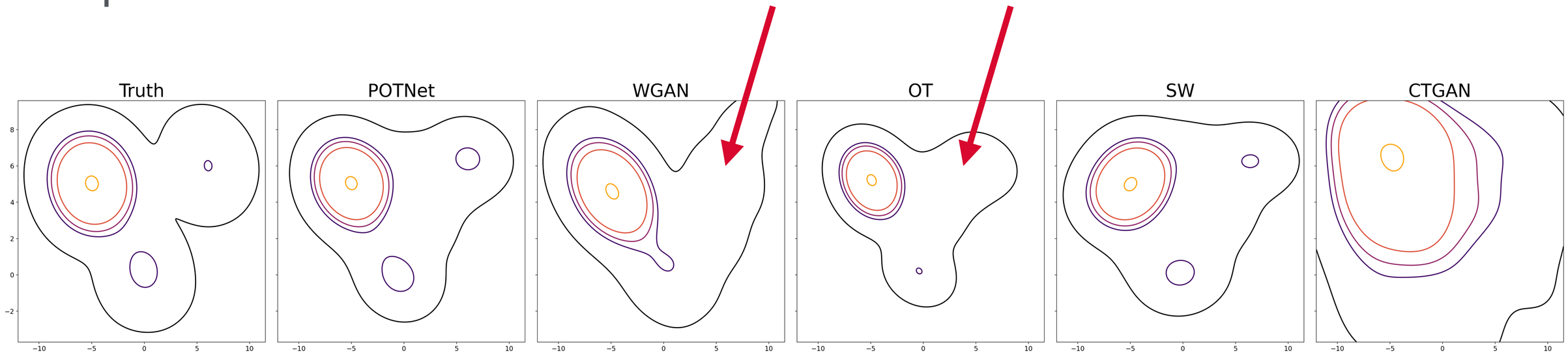
$$\mathcal{D}(\mu,\nu) := \underbrace{W_1(\mu,\nu)}_{joint\ distance} + \sum_{j=1}^{d} \underbrace{\lambda_j W_1((p_j)_*\mu, (p_j)_*\nu)}_{marginal\ penalty}$$

regularization factor on the $j$th marginal

- Use the marginal distribution to **constrain** possible mass allocations and use the joint loss to **identify** the correct mass allocation.

# The POTNet Framework

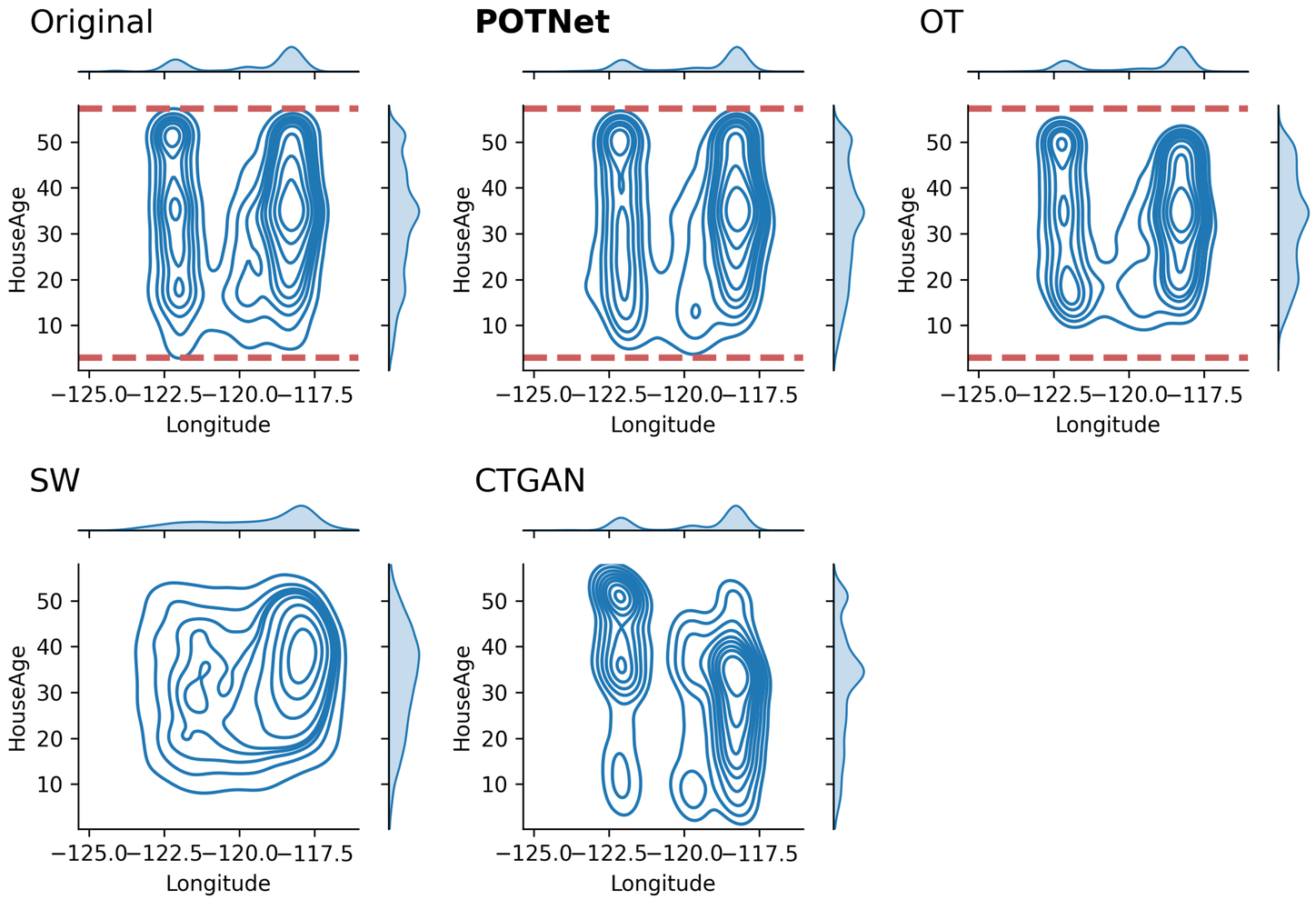
## Empirical Evaluations




20 dimensional mixture of Gaussian example

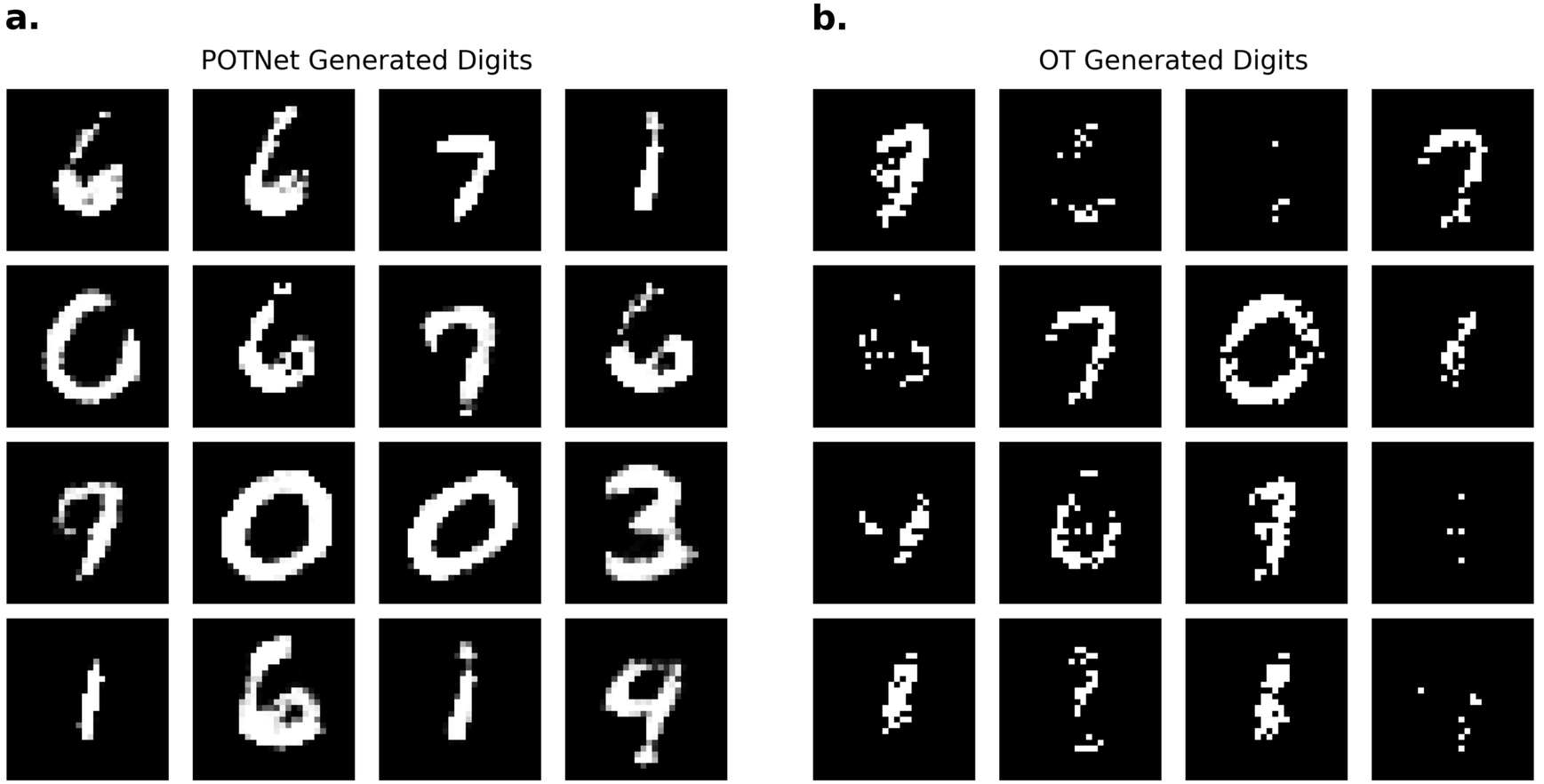
# The POTNet Framework

## Real Data Examples: California Housing Data



# The POTNet Framework

## Real Data Examples: MNIST Digit Generation




Using feedforward neural network architecture

# Potential connections to LassoNet…?

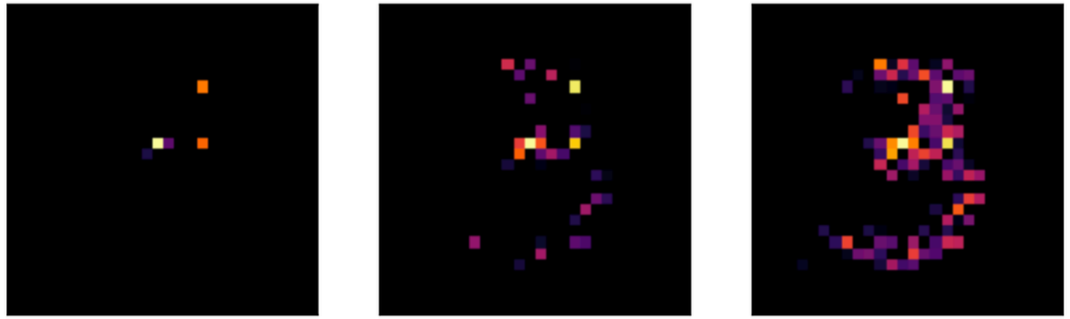


*Figure 7.* Results for LassoNet in choosing the most informative pixels of images of the digit 3 in the MNIST dataset, for three different penalty levels ($\lambda = 5, \lambda = 1, \lambda = 0.1$).

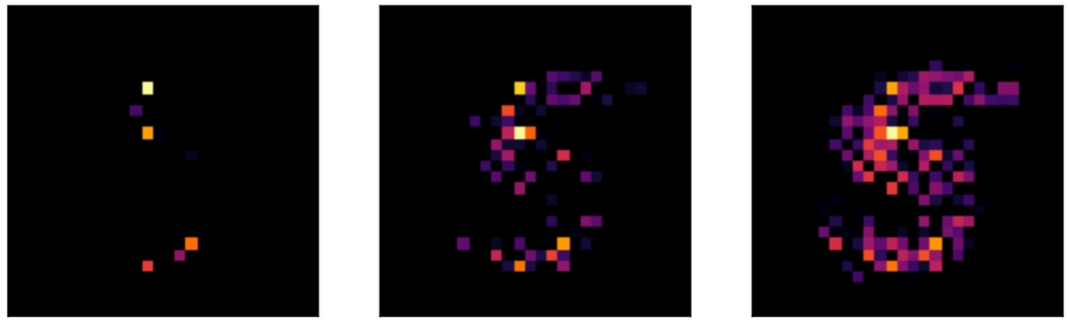


*Figure 8.* Results for LassoNet in choosing the most informative pixels of images of the digit 5 in the MNIST dataset, for the three penalty levels.