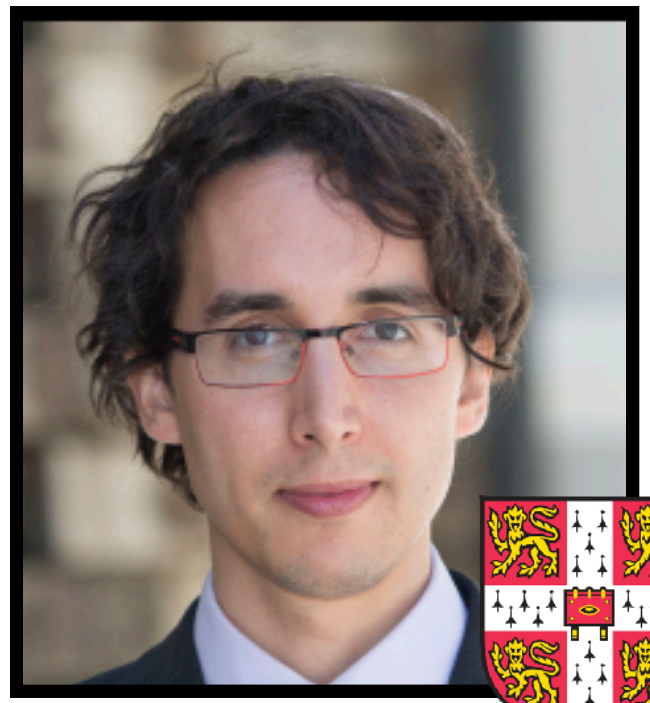
Dropout as a Structured Shrinkage Prior

Eric Nalisnick, **José Miguel Hernández-Lobato**, **Padhraic Smyth**

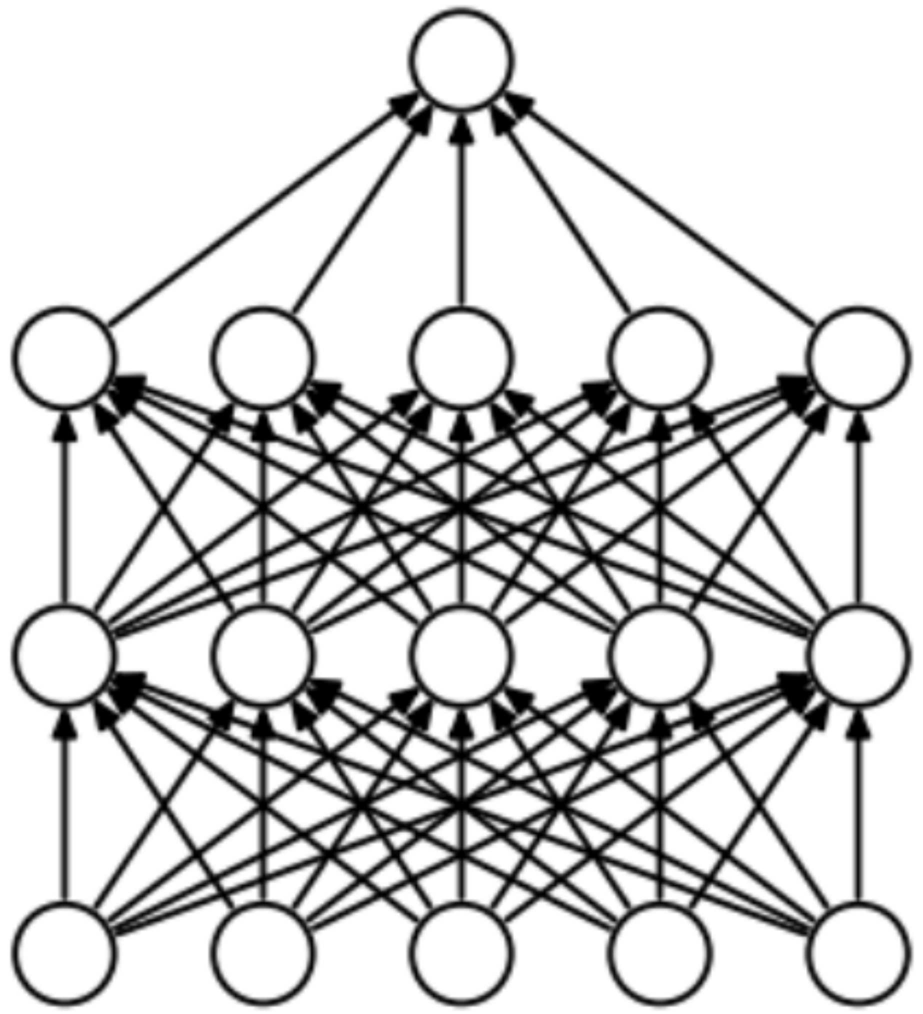


University of Cambridge   University of California, Irvine

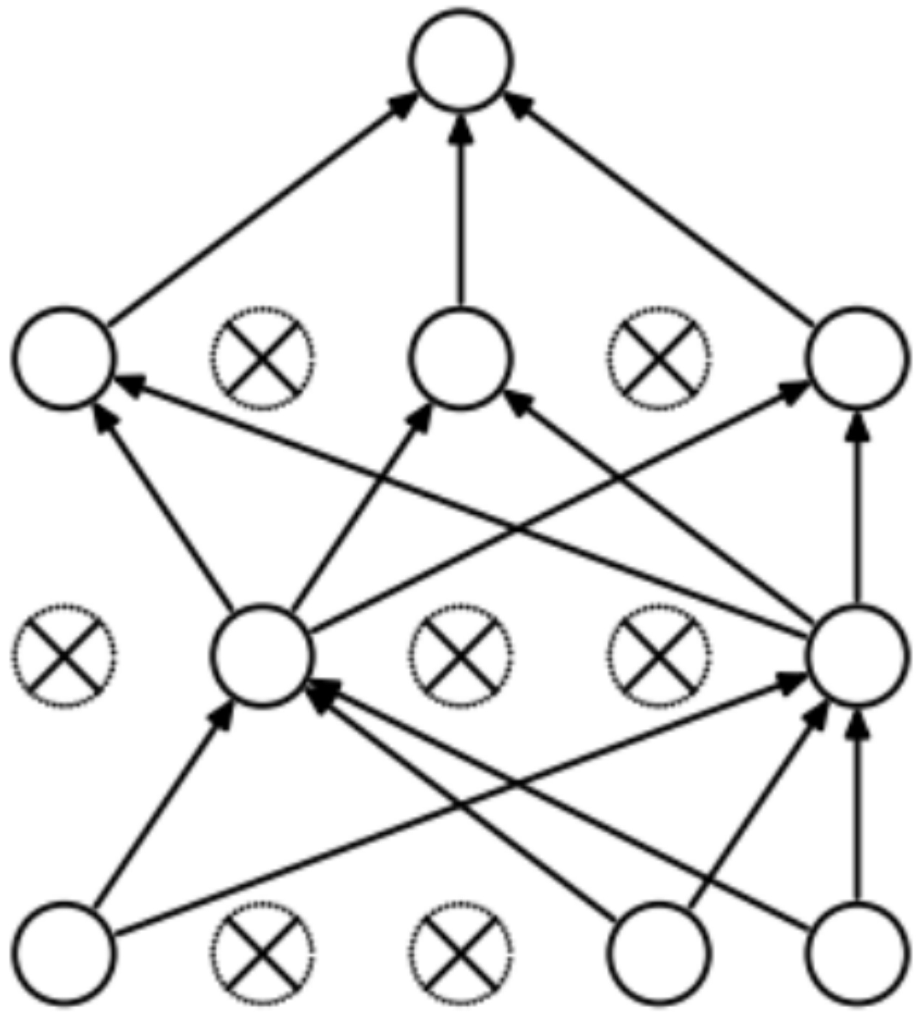
Dropout & Multiplicative Noise

Improving neural networks by preventing co-adaptation of feature detectors (2012)

G. E. Hinton*, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov
Department of Computer Science, University of Toronto,
6 King's College Rd, Toronto, Ontario M5S 3G4, Canada



Standard Neural Network

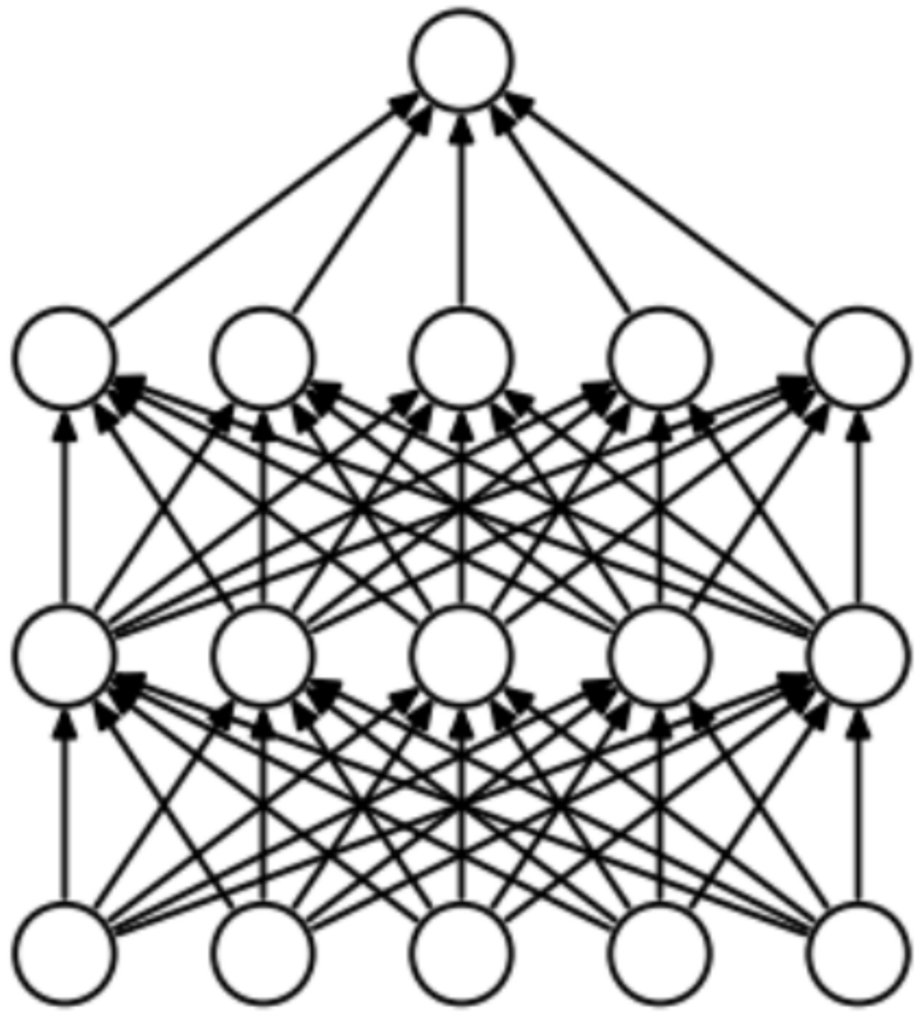


After Applying Dropout

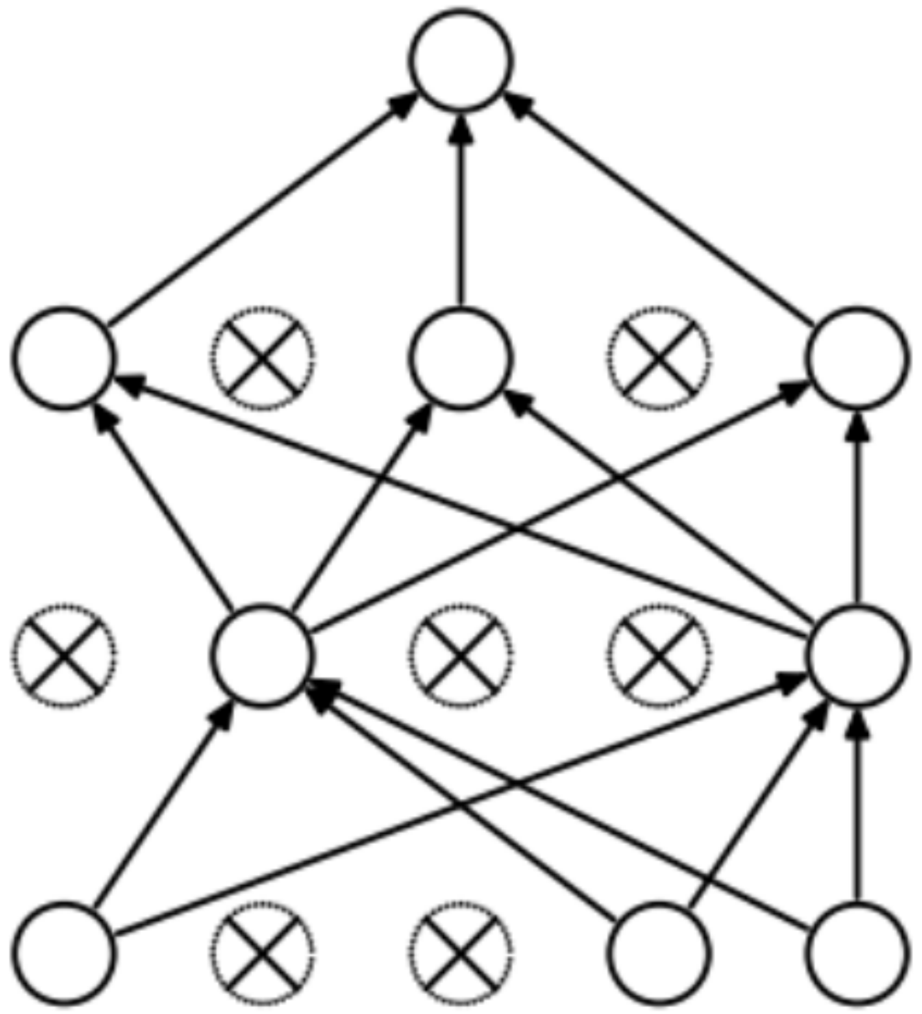
Dropout & Multiplicative Noise

Improving neural networks by preventing co-adaptation of feature detectors (2012)

G. E. Hinton*, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov
Department of Computer Science, University of Toronto,
6 King's College Rd, Toronto, Ontario M5S 3G4, Canada



Standard Neural Network



After Applying Dropout

Implementation as **Multiplicative Noise**:

$$\mathbf{h}_{n,l} = f_l(\mathbf{h}_{n,l-1}\mathbf{\Lambda}_l\mathbf{W}_l)$$

Hidden Units

Weights

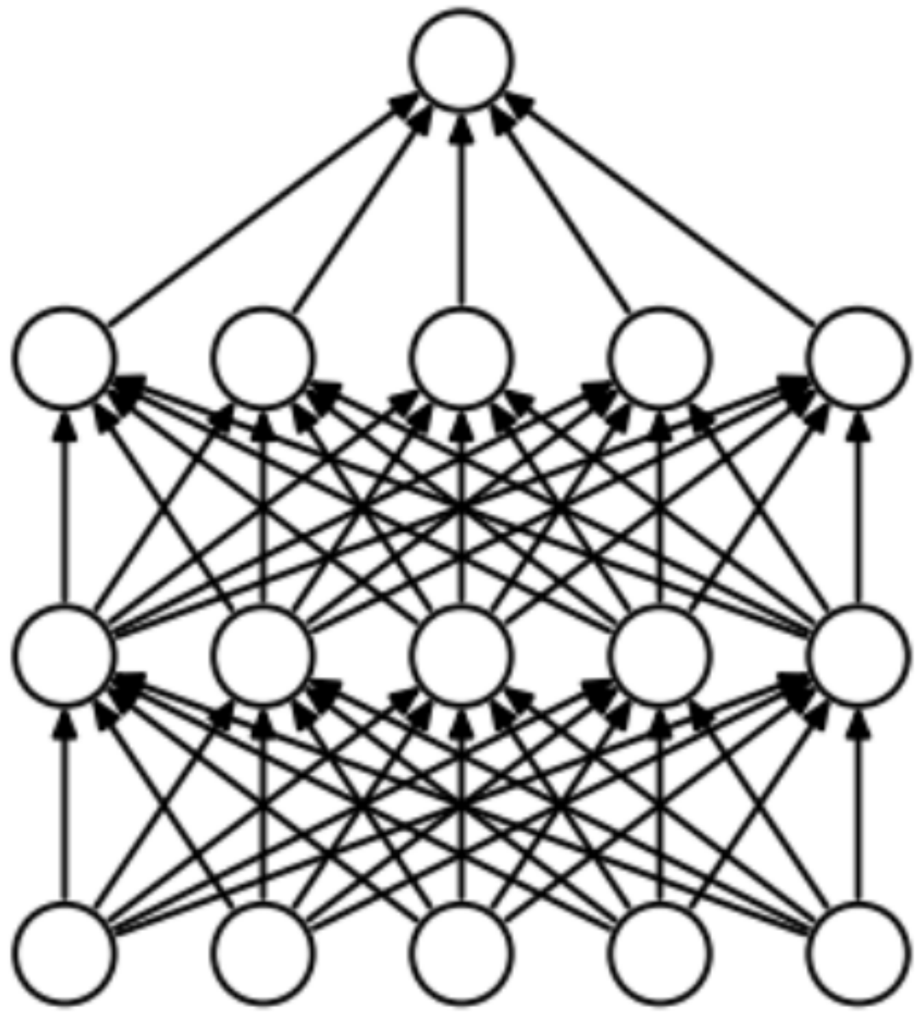
Diagonal Matrix of Random Variables

$$\lambda_{i,i} \sim p(\lambda)$$

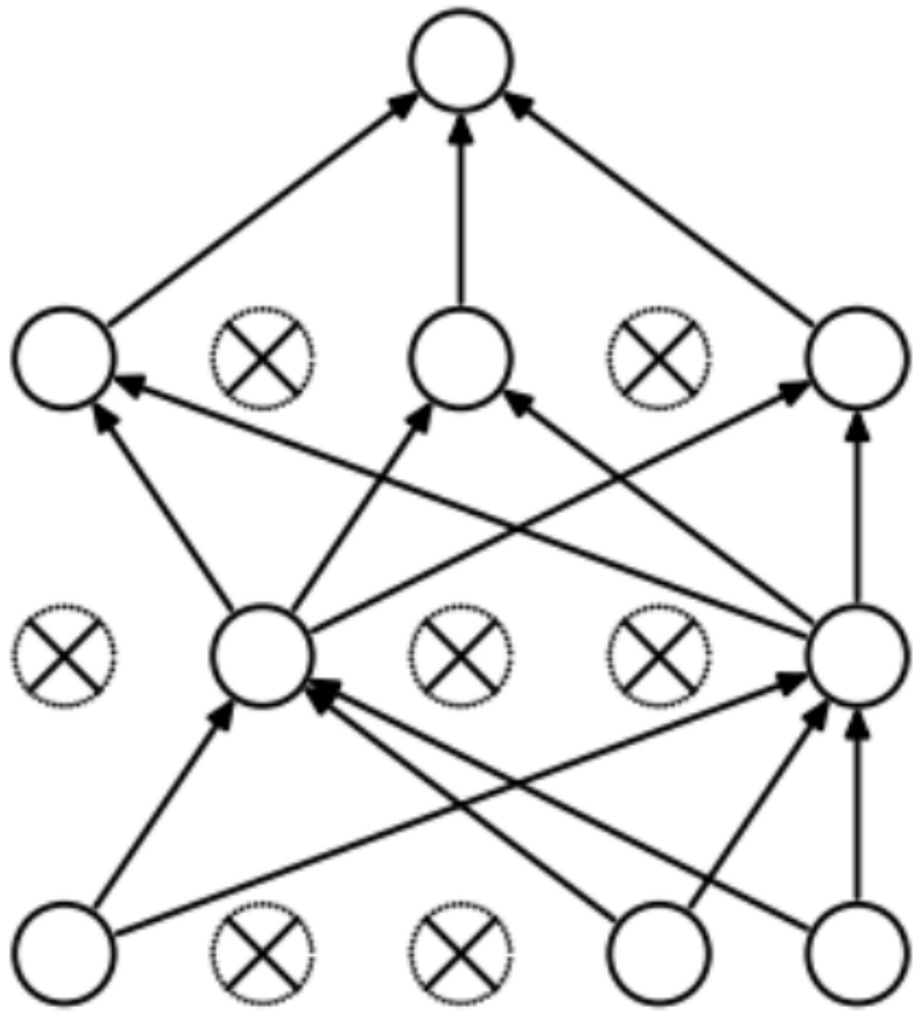
Dropout & Multiplicative Noise

Improving neural networks by preventing co-adaptation of feature detectors (2012)

G. E. Hinton*, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov
Department of Computer Science, University of Toronto,
6 King's College Rd, Toronto, Ontario M5S 3G4, Canada



Standard Neural Network



After Applying Dropout

Implementation as **Multiplicative Noise**:

$$\mathbf{h}_{n,l} = f_l(\mathbf{h}_{n,l-1}\mathbf{\Lambda}_l\mathbf{W}_l)$$

Hidden Units

Weights

Diagonal Matrix of Random Variables

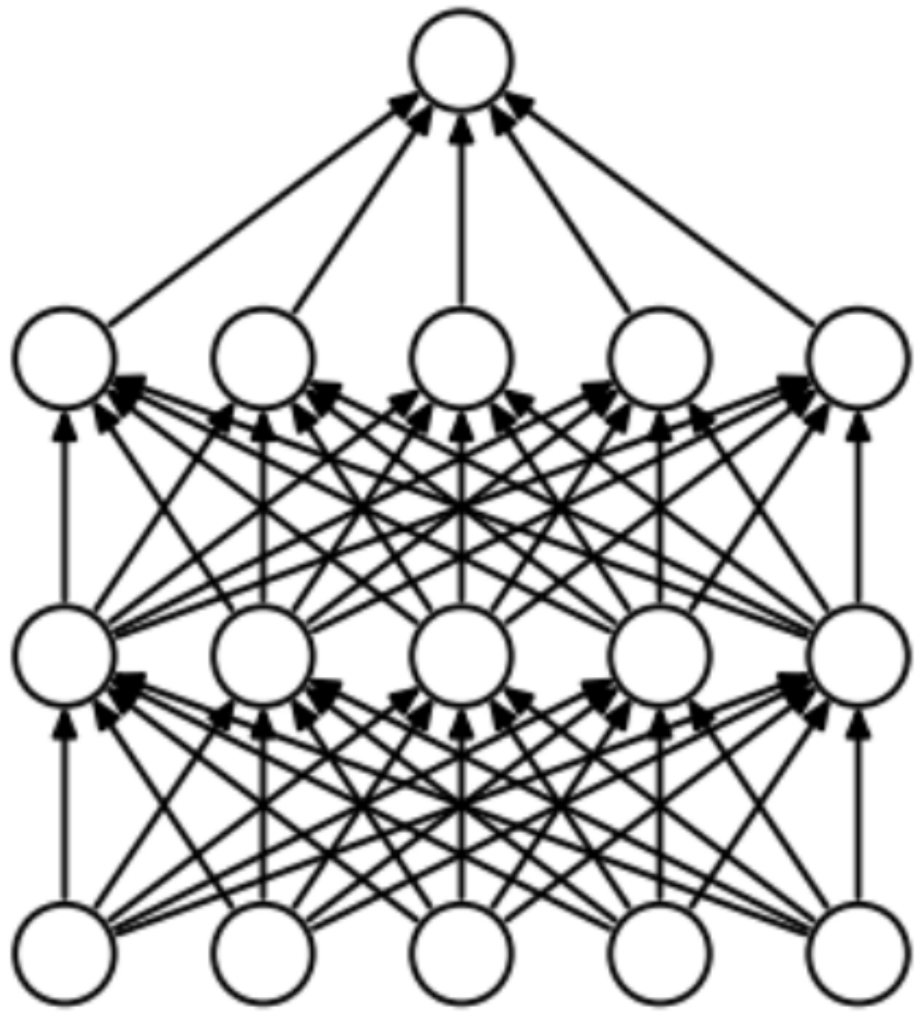
$$\lambda_{i,i} \sim p(\lambda)$$

- Dropout corresponds to **p(λ)** being Bernoulli.

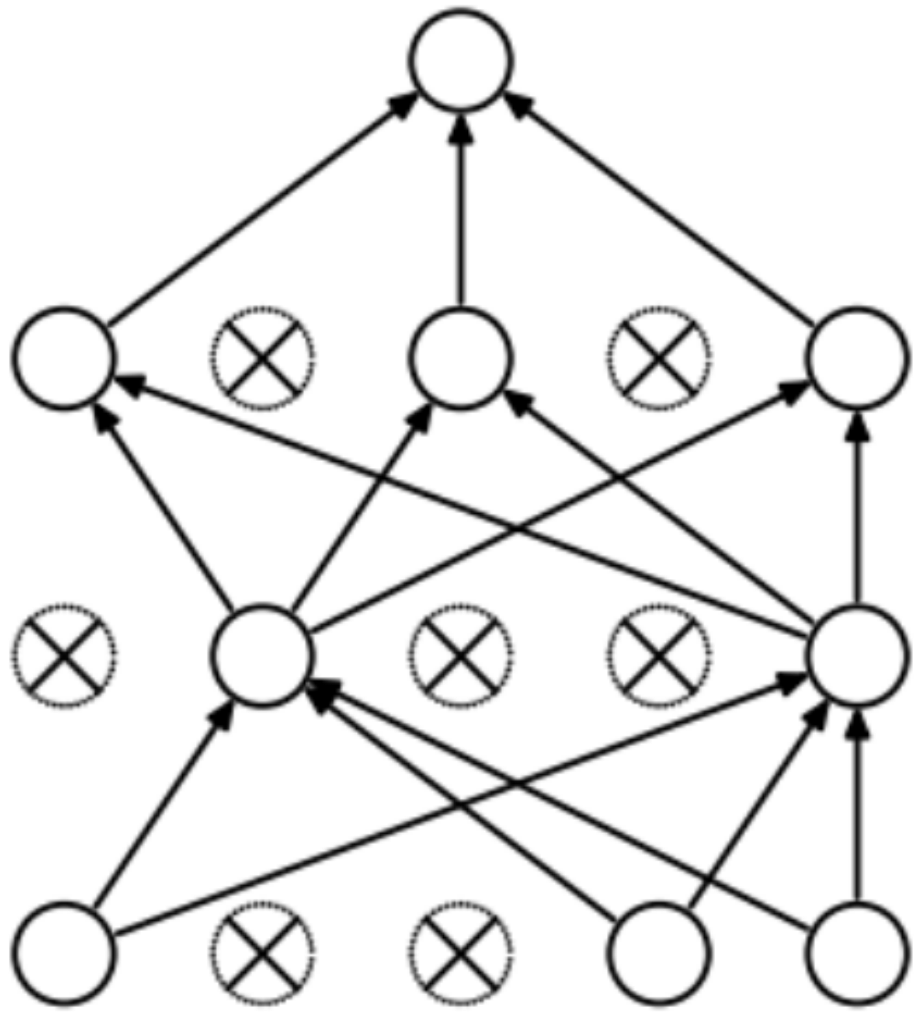
Dropout & Multiplicative Noise

Improving neural networks by preventing co-adaptation of feature detectors (2012)

G. E. Hinton*, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov
Department of Computer Science, University of Toronto,
6 King's College Rd, Toronto, Ontario M5S 3G4, Canada



Standard Neural Network



After Applying Dropout

Implementation as **Multiplicative Noise**:

$$\mathbf{h}_{n,l} = f_l(\mathbf{h}_{n,l-1}\mathbf{\Lambda}_l\mathbf{W}_l)$$

Hidden Units

Weights

Diagonal Matrix of Random Variables

$$\lambda_{i,i} \sim p(\lambda)$$

- Dropout corresponds to **p(λ)** being Bernoulli.
- Gaussian, beta, and uniform noise have been shown to work as well.

Dropout as a Gaussian Scale Mixture

Dropout as a Gaussian Scale Mixture

Gaussian Scale Mixtures

A random variable θ is a **Gaussian scale mixture** *iff* it can be expressed as the product of a Gaussian random variable and an independent scalar random variable [Beale & Mallows, 1959]:

$$\theta \stackrel{d}{=} \alpha z, \quad z \sim \mathrm{N}(0, \sigma_0^2), \;\; \alpha \sim p(\alpha)$$

Dropout as a Gaussian Scale Mixture

Gaussian Scale Mixtures

A random variable θ is a **Gaussian scale mixture** *iff* it can be expressed as the product of a Gaussian random variable and an independent scalar random variable [Beale & Mallows, 1959]:

$$\theta \stackrel{d}{=} \alpha z, \quad z \sim \mathrm{N}(0, \sigma_0^2), \;\; \alpha \sim p(\alpha)$$

Can be reparametrized into a **hierarchical form**:

$$z \sim \mathrm{N}(0, \alpha^2 \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Dropout as a Gaussian Scale Mixture

Gaussian Scale Mixtures

A random variable θ is a **Gaussian scale mixture** *iff* it can be expressed as the product of a Gaussian random variable and an independent scalar random variable [Beale & Mallows, 1959]:

$$\theta \stackrel{d}{=} \alpha z, \quad z \sim \mathrm{N}(0, \sigma_0^2), \;\; \alpha \sim p(\alpha)$$

Can be reparametrized into a **hierarchical form**:

$$z \sim \mathrm{N}(0, \alpha^2 \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Let's assume a **Gaussian prior on the NN weights**…

$$f_l(\mathbf{h}_{n,l-1} \mathbf{\Lambda}_l \mathbf{W}_l)$$

Noise

$$\lambda_{i,i} \sim p(\lambda)$$

Weights

$$w_{i,j} \sim \mathrm{N}(0, \sigma_0^2)$$

Dropout as a Gaussian Scale Mixture

Gaussian Scale Mixtures

A random variable θ is a **Gaussian scale mixture** *iff* it can be expressed as the product of a Gaussian random variable and an independent scalar random variable [Beale & Mallows, 1959]:

$$\theta \stackrel{d}{=} \alpha z, \quad z \sim \mathrm{N}(0, \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Can be reparametrized into a **hierarchical form**:

$$z \sim \mathrm{N}(0, \alpha^2 \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Let's assume a **Gaussian prior on the NN weights**…

$$f_l(\mathbf{h}_{n,l-1} \underbrace{\mathbf{\Lambda}_l \mathbf{W}_l})$$

Definition of a Gaussian Scale Mixture

Dropout as a Gaussian Scale Mixture

Gaussian Scale Mixtures

A random variable θ is a **Gaussian scale mixture** *iff* it can be expressed as the product of a Gaussian random variable and an independent scalar random variable [Beale & Mallows, 1959]:

$$\theta \stackrel{d}{=} \alpha z, \quad z \sim \mathrm{N}(0, \sigma_0^2), \;\; \alpha \sim p(\alpha)$$

Can be reparametrized into a **hierarchical form**:

$$z \sim \mathrm{N}(0, \alpha^2 \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Let's assume a **Gaussian prior on the NN weights**…

$$f_l(\mathbf{h}_{n,l-1} \underbrace{\mathbf{\Lambda}_l \mathbf{W}_l})$$

Definition of a Gaussian Scale Mixture

SWITCH TO HIERARCHICAL PARAMETRIZATION ↓

Dropout as a Gaussian Scale Mixture

Gaussian Scale Mixtures

A random variable θ is a **Gaussian scale mixture** *iff* it can be expressed as the product of a Gaussian random variable and an independent scalar random variable [Beale & Mallows, 1959]:

$$\theta \stackrel{d}{=} \alpha z, \quad z \sim \mathrm{N}(0, \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Can be reparametrized into a **hierarchical form**:

$$z \sim \mathrm{N}(0, \alpha^2 \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Let's assume a **Gaussian prior on the NN weights**…

$$f_l(\mathbf{h}_{n,l-1} \boldsymbol{\Lambda}_l \mathbf{W}_l)$$

Definition of a Gaussian Scale Mixture

SWITCH TO HIERARCHICAL PARAMETRIZATION

$$f_l(\mathbf{h}_{n,l-1} \mathbf{W}_l)$$

$$w_{i,j} \sim \mathrm{N}(0, \lambda_{i,i}^2 \sigma_0^2)$$

Noise distribution becomes a scale prior

Dropout as a Gaussian Scale Mixture

Can translate noise distributions into the marginal prior they induce on the NN weights…

Noise Model $p(\lambda)$	Variance Prior $p(\lambda^2)$	Marginal Prior $p(w)$
Bernoulli	Bernoulli	Spike-and-Slab
Gaussian	χ^2	Generalized Hyperbolic
Rayleigh	Exponential	Laplace
Inverse Nakagami	Γ^{-1}	Student-t
Half-Cauchy	Unnamed	Horseshoe

Dropout's Scale Structure

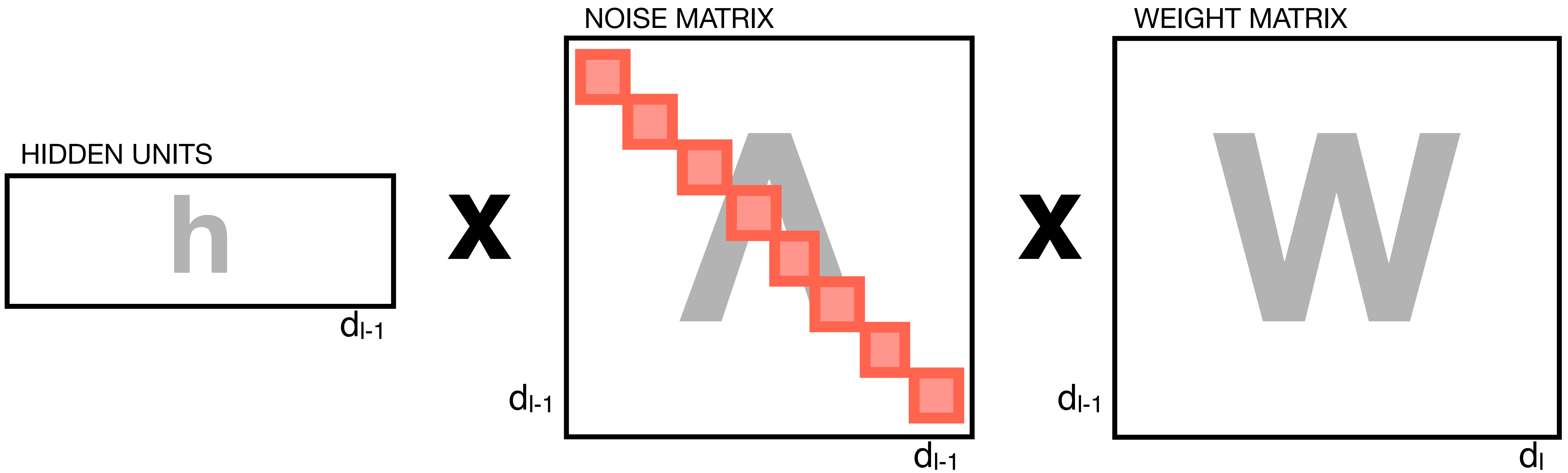
Sampling noise for each hidden unit induces a particular structure…

$$f_l(\mathbf{h}_{n,l-1}\mathbf{\Lambda}_l\mathbf{W}_l) \qquad w_{i,j} \sim \mathrm{N}(0, \sigma_0^2)$$

Dropout's Scale Structure

Sampling noise for each hidden unit induces a particular structure…

$$f_l(\mathbf{h}_{n,l-1}\mathbf{\Lambda}_l\mathbf{W}_l) \qquad w_{i,j} \sim \mathrm{N}(0, \sigma_0^2)$$

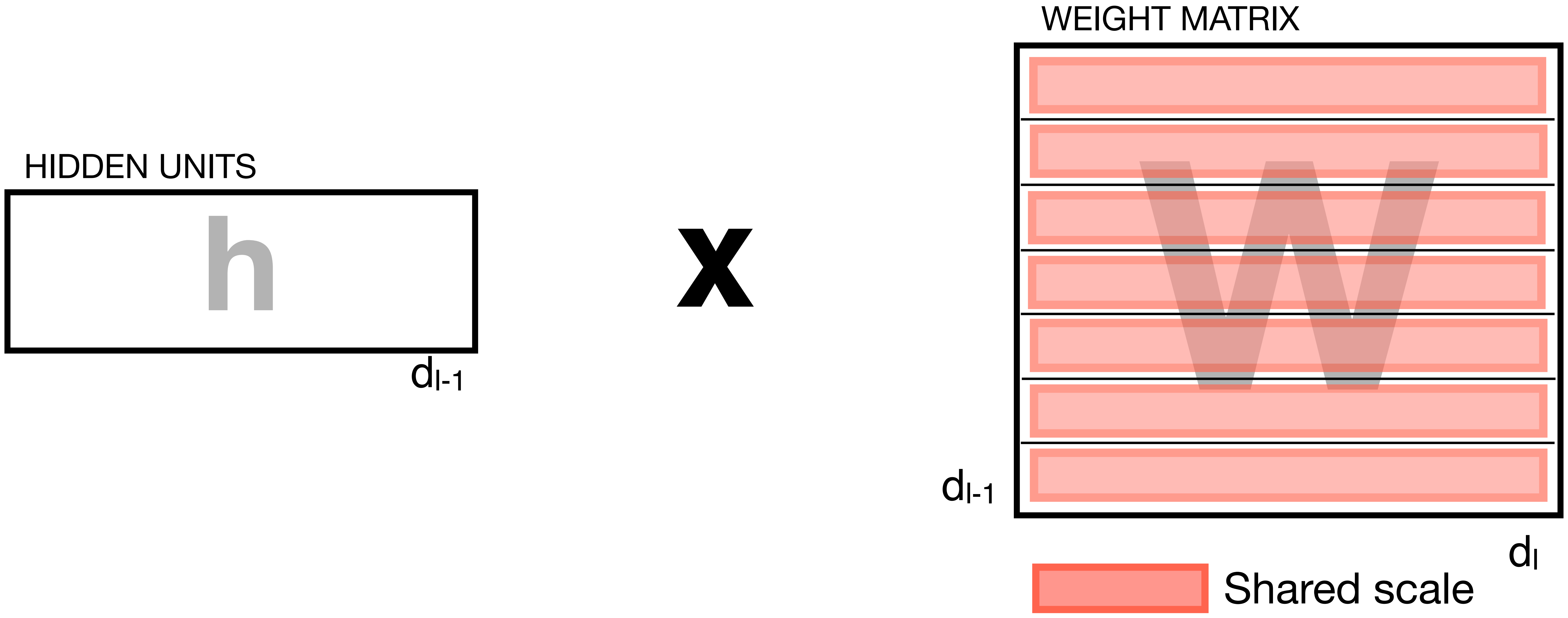


Dropout's Scale Structure

Sampling noise for each hidden unit induces a particular structure…

$$f_l(\mathbf{h}_{n,l-1}\mathbf{W}_l) \qquad w_{i,j} \sim \mathrm{N}(0, \lambda_{i,i}^2 \sigma_0^2)$$

i indexes rows



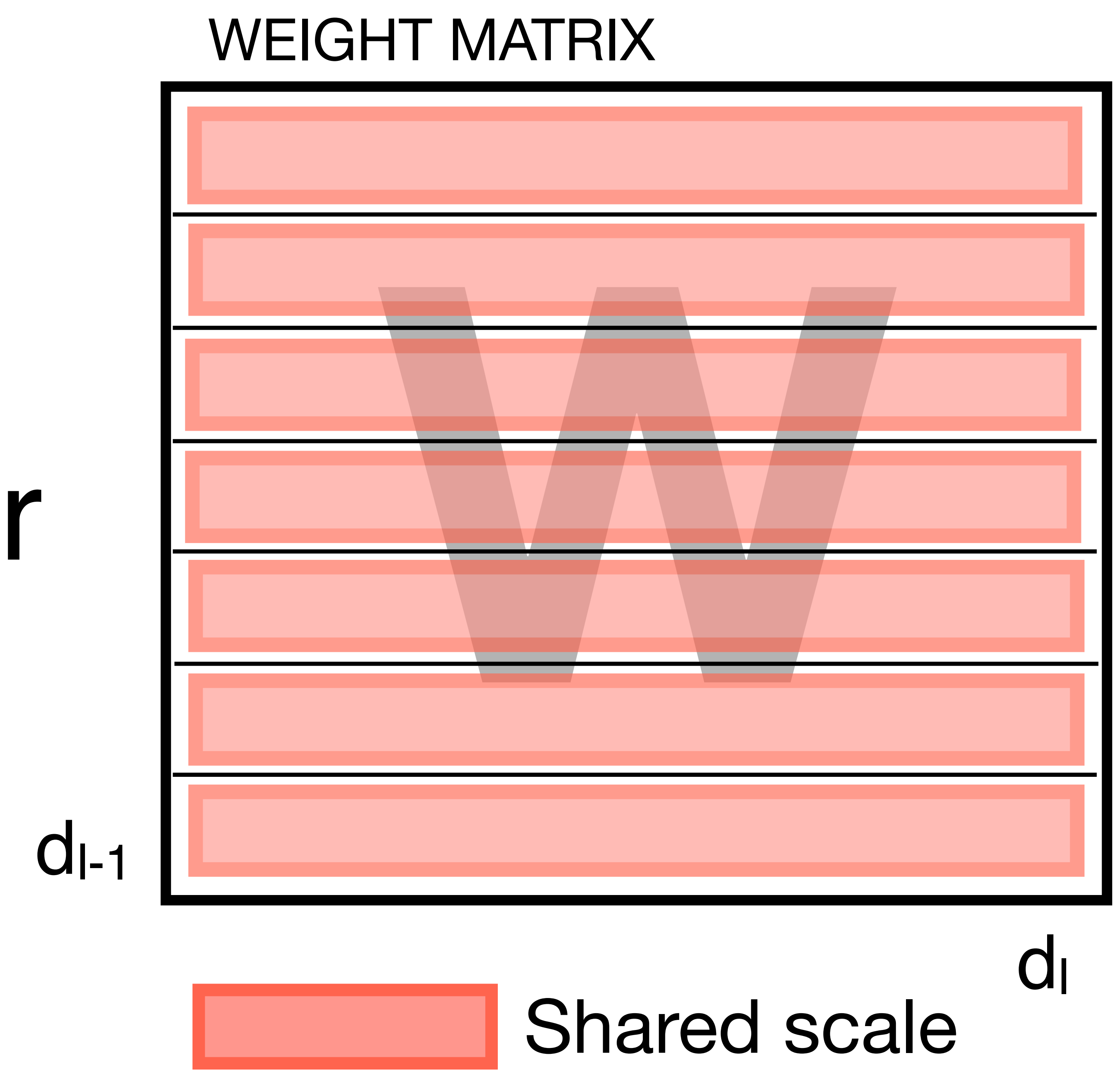
Dropout's Scale Structure

Sampling noise for each hidden unit induces a particular structure…

$$f_l(\mathbf{h}_{n,l-1}\mathbf{W}_l) \qquad w_{i,j} \sim \mathrm{N}(0, \lambda_{i,i}^2 \sigma_0^2)$$

i indexes rows

Same structure as the **automatic relevance determination (ARD)** prior proposed by D. MacKay and R. Neal for Bayesian NNs (1994).



Summary

- Under mild assumptions, **multiplicative noise is equivalent to a Gauss. scale mixture prior with ARD structure**.

Summary

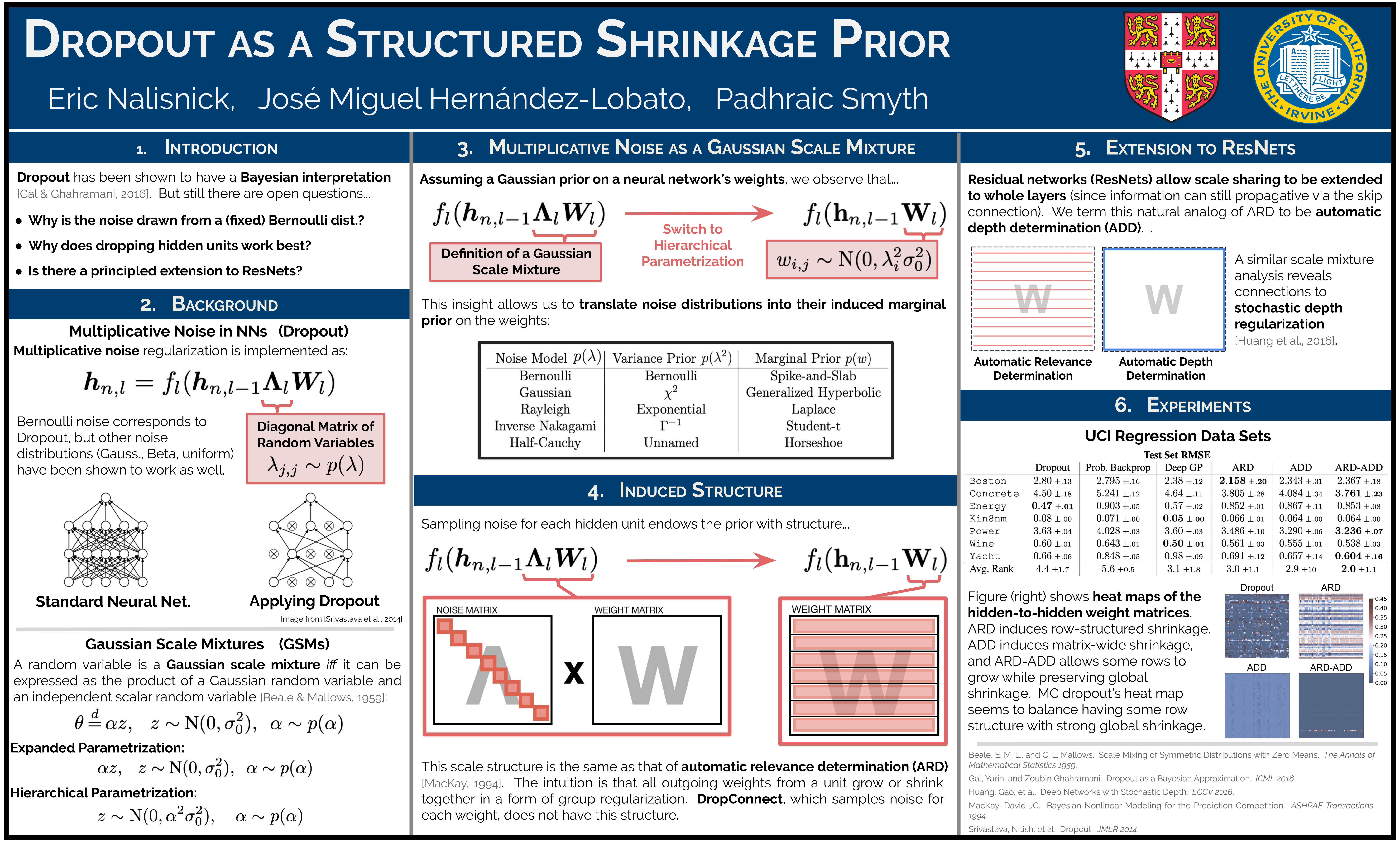
- Under mild assumptions, **multiplicative noise is equivalent to a Gauss. scale mixture prior with ARD structure**.

- This **decouples dropout's Bayesian interpretation from variational inference**, allowing for any inference strategy.

Summary

- Under mild assumptions, **multiplicative noise is equivalent to a Gauss. scale mixture prior with ARD structure**.
- This **decouples dropout's Bayesian interpretation from variational inference**, allowing for any inference strategy.
- Provides a **'recipe' for translating noise distributions into priors**, better revealing their modeling assumptions.

For more details, please visit our poster (#84)



DROPOUT AS A STRUCTURED SHRINKAGE PRIOR

Eric Nalisnick, José Miguel Hernández-Lobato, Padhraic Smyth

1. INTRODUCTION

Dropout has been shown to have a **Bayesian interpretation** [Gal & Ghahramani, 2016]. But still there are open questions...

- **Why is the noise drawn from a (fixed) Bernoulli dist.?**
- **Why does dropping hidden units work best?**
- **Is there a principled extension to ResNets?**

2. BACKGROUND

Multiplicative Noise in NNs (Dropout)

Multiplicative noise regularization is implemented as:

$$\boldsymbol{h}_{n,l} = f_l(\boldsymbol{h}_{n,l-1}\boldsymbol{\Lambda}_l\boldsymbol{W}_l)$$

Bernoulli noise corresponds to Dropout, but other noise distributions (Gauss., Beta, uniform) have been shown to work as well.

Diagonal Matrix of Random Variables $\lambda_{j,j} \sim p(\lambda)$

Standard Neural Net. **Applying Dropout**

Image from [Srivastava et al., 2014]

Gaussian Scale Mixtures (GSMs)

A random variable is a **Gaussian scale mixture** *iff* it can be expressed as the product of a Gaussian random variable and an independent scalar random variable [Beale & Mallows, 1959]:

$$\theta \stackrel{d}{=} \alpha z, \quad z \sim \text{N}(0, \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Expanded Parametrization:

$$\alpha z, \quad z \sim \text{N}(0, \sigma_0^2), \quad \alpha \sim p(\alpha)$$

Hierarchical Parametrization:

$$z \sim \text{N}(0, \alpha^2\sigma_0^2), \quad \alpha \sim p(\alpha)$$

3. MULTIPLICATIVE NOISE AS A GAUSSIAN SCALE MIXTURE

Assuming a Gaussian prior on a neural network's weights, we observe that...

$$f_l(\boldsymbol{h}_{n,l-1}\boldsymbol{\Lambda}_l\boldsymbol{W}_l) \quad \rightarrow \quad f_l(\mathbf{h}_{n,l-1}\mathbf{W}_l)$$

Definition of a Gaussian Scale Mixture — Switch to Hierarchical Parametrization — $w_{i,j} \sim \text{N}(0, \lambda_i^2\sigma_0^2)$

This insight allows us to **translate noise distributions into their induced marginal prior** on the weights:

Noise Model $p(\lambda)$	Variance Prior $p(\lambda^2)$	Marginal Prior $p(w)$
Bernoulli	Bernoulli	Spike-and-Slab
Gaussian	χ^2	Generalized Hyperbolic
Rayleigh	Exponential	Laplace
Inverse Nakagami	Γ^{-1}	Student-t
Half-Cauchy	Unnamed	Horseshoe

4. INDUCED STRUCTURE

Sampling noise for each hidden unit endows the prior with structure...

$$f_l(\boldsymbol{h}_{n,l-1}\boldsymbol{\Lambda}_l\boldsymbol{W}_l) \quad \rightarrow \quad f_l(\mathbf{h}_{n,l-1}\mathbf{W}_l)$$

NOISE MATRIX Λ x WEIGHT MATRIX W — WEIGHT MATRIX W

This scale structure is the same as that of **automatic relevance determination (ARD)** [MacKay, 1994]. The intuition is that all outgoing weights from a unit grow or shrink together in a form of group regularization. **DropConnect**, which samples noise for each weight, does not have this structure.

5. EXTENSION TO RESNETS

Residual networks (ResNets) allow scale sharing to be extended to whole layers (since information can still propagative via the skip connection). We term this natural analog of ARD to be **automatic depth determination (ADD)**. .

W — **Automatic Relevance Determination**

W — **Automatic Depth Determination**

A similar scale mixture analysis reveals connections to **stochastic depth regularization** [Huang et al., 2016].

6. EXPERIMENTS

UCI Regression Data Sets

Test Set RMSE

	Dropout	Prob. Backprop	Deep GP	ARD	ADD	ARD-ADD
Boston	2.80 ±.13	2.795 ±.16	2.38 ±.12	**2.158 ±.20**	2.343 ±.31	2.367 ±.18
Concrete	4.50 ±.18	5.241 ±.12	4.64 ±.11	3.805 ±.28	4.084 ±.34	**3.761 ±.23**
Energy	**0.47 ±.01**	0.903 ±.05	0.57 ±.02	0.852 ±.01	0.867 ±.11	0.853 ±.08
Kin8nm	0.08 ±.00	0.071 ±.00	**0.05 ±.00**	0.066 ±.01	0.064 ±.00	0.064 ±.00
Power	3.63 ±.04	4.028 ±.03	3.60 ±.03	3.486 ±.10	3.290 ±.06	**3.236 ±.07**
Wine	0.60 ±.01	0.643 ±.01	**0.50 ±.01**	0.561 ±.03	0.555 ±.01	0.538 ±.03
Yacht	0.66 ±.06	0.848 ±.05	0.98 ±.09	0.691 ±.12	0.657 ±.14	**0.604 ±.16**
Avg. Rank	4.4 ±1.7	5.6 ±0.5	3.1 ±1.8	3.0 ±1.1	2.9 ±10	**2.0 ±1.1**

Figure (right) shows **heat maps of the hidden-to-hidden weight matrices**. ARD induces row-structured shrinkage, ADD induces matrix-wide shrinkage, and ARD-ADD allows some rows to grow while preserving global shrinkage. MC dropout's heat map seems to balance having some row structure with strong global shrinkage.

Dropout ARD ADD ARD-ADD

0.45 0.40 0.35 0.30 0.25 0.20 0.15 0.10 0.05 0.00

Beale, E. M. L., and C. L. Mallows. Scale Mixing of Symmetric Distributions with Zero Means. *The Annals of Mathematical Statistics 1959*.

Gal, Yarin, and Zoubin Ghahramani. Dropout as a Bayesian Approximation. *ICML 2016*.

Huang, Gao, et al. Deep Networks with Stochastic Depth. *ECCV 2016*.

MacKay, David JC. Bayesian Nonlinear Modeling for the Prediction Competition. *ASHRAE Transactions 1994*.

Srivastava, Nitish, et al. Dropout. *JMLR 2014*.