The Functional Neural Process (FNP) & The Functional Process VAE (FP-VAE)

Max Welling

Co-authors: Christos Louizos, Xiahan Shi, Klamer Schutte

First author

AMLAB Amsterdam Machine Learning Lab

University of Amsterdam

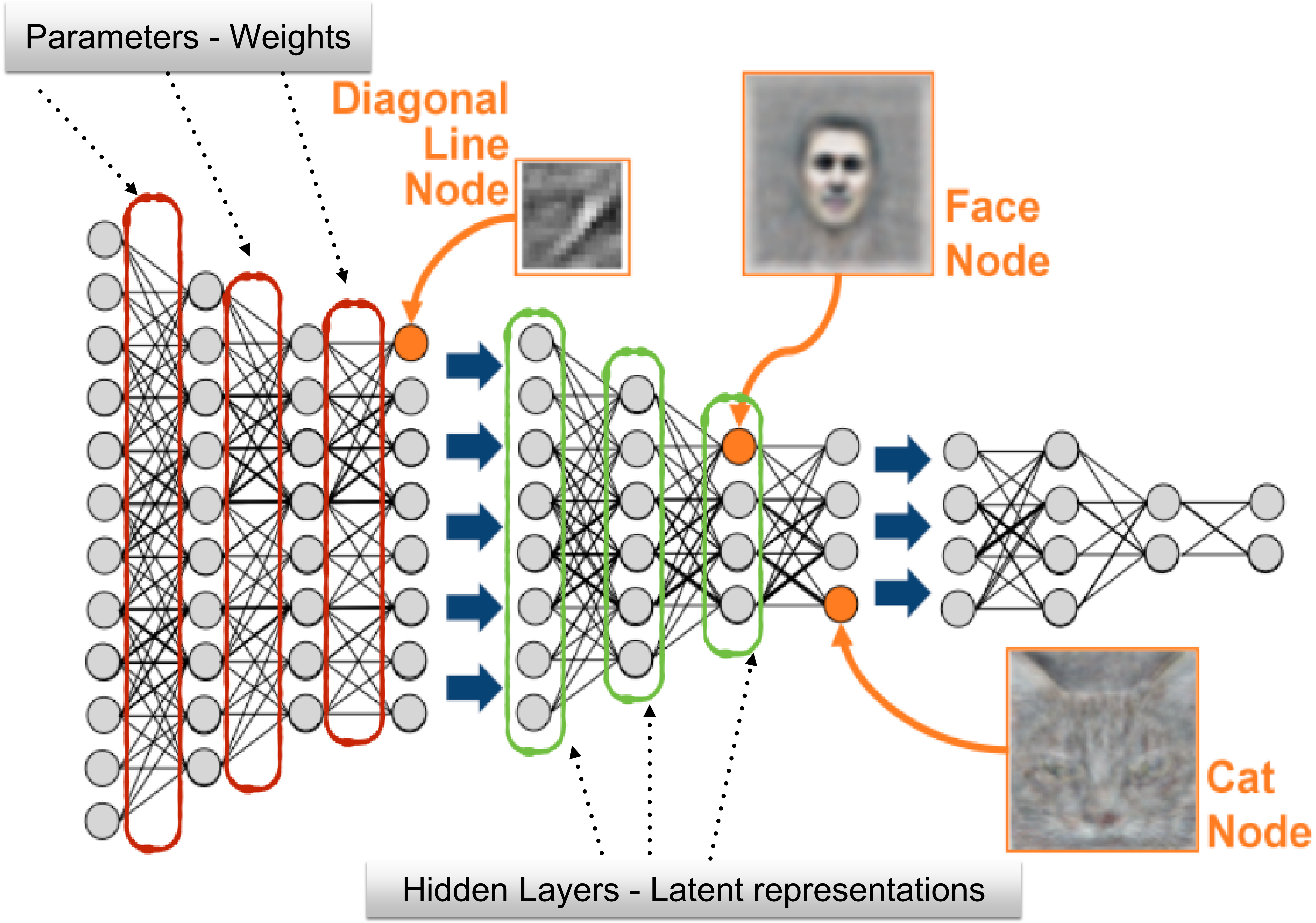
BOSCH Invented for life

TNO

What is this talk about?

- I want to do deep learning
- I want to be Bayesian
- I don’t want to place priors on a million parameters
- I want to work with “big data”
- (I want to do cool mathematics)
- *What should I do?*

Motivation



Original image from: http://scyfer.nl/wp-content/uploads/2014/05/Deep_Neural_Network.png

- Deep neural networks excel at various **predictive tasks**
- But for decision making you also need **confidence levels**
- Detect that we have not seen this before (**epistemic uncertainty**)

Bayesian Neural Networks (BNNs)

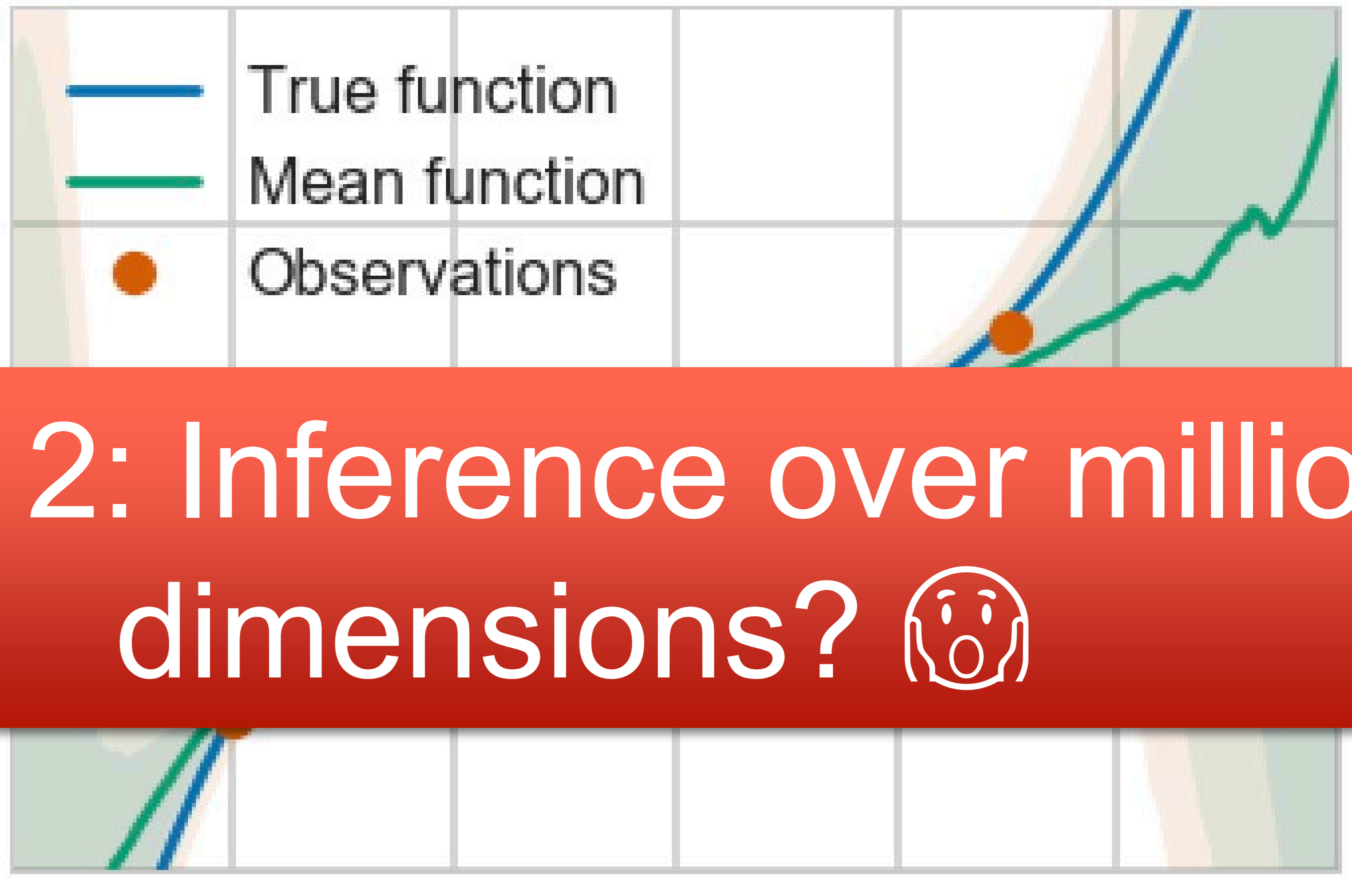
- Allows neural networks to **estimate uncertainty**

Problem 1: what is an appropriate choice of p(**w**)? 😱

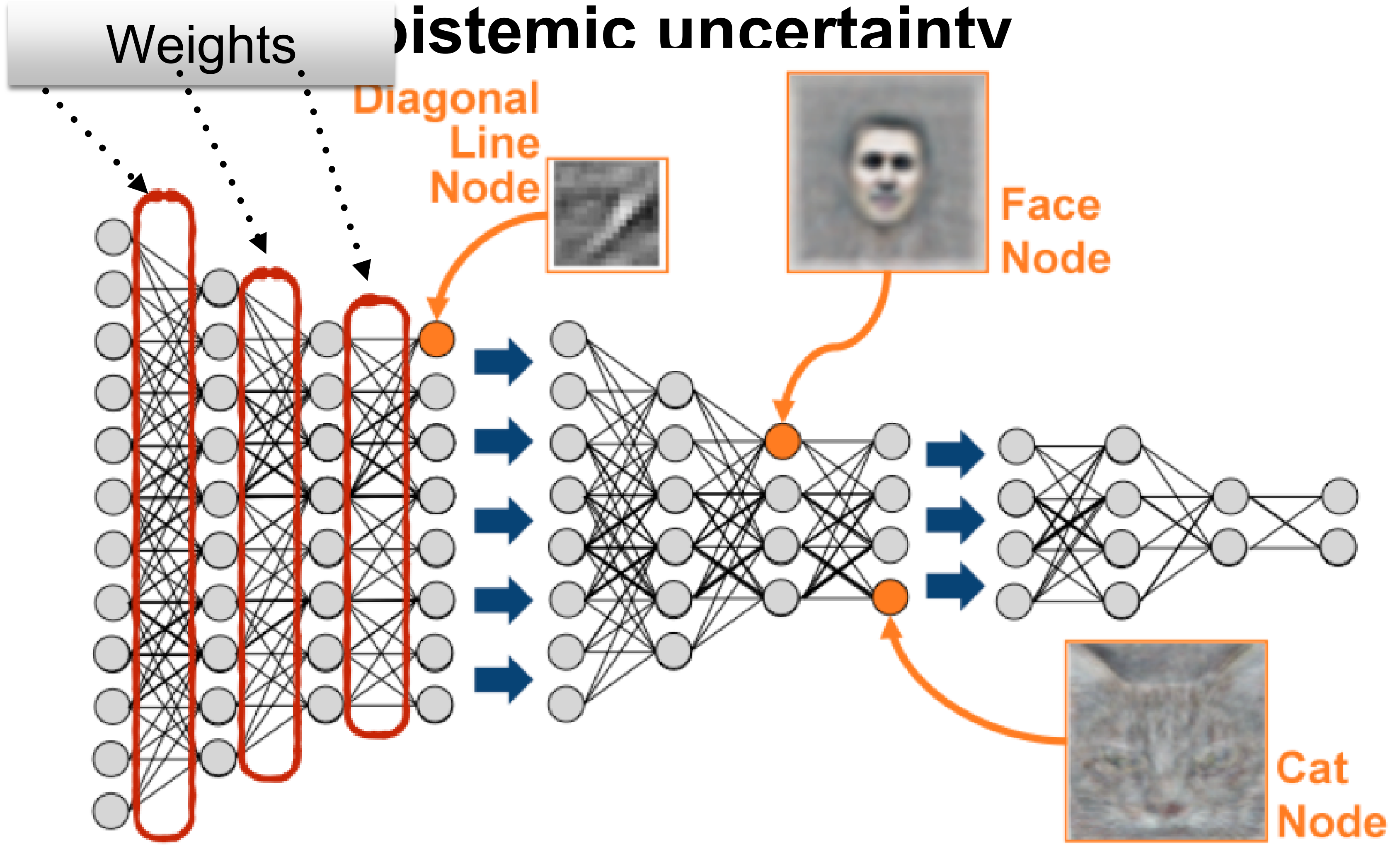
- p
  over the weights **w** of the network

- *Summarize the uncertainty in the* **posterior distribution** over **w**

$$p(\mathbf{w}|x_{1:N}, y_{1:N}) = \frac{p(\mathbf{w})\prod_{i=1}^{N} p(y_i|x_i,\mathbf{w})}{\int p(\mathbf{w})\prod_{i=1}^{N} p(y_i|x_i,\mathbf{w})\mathrm{d}\mathbf{w}}$$

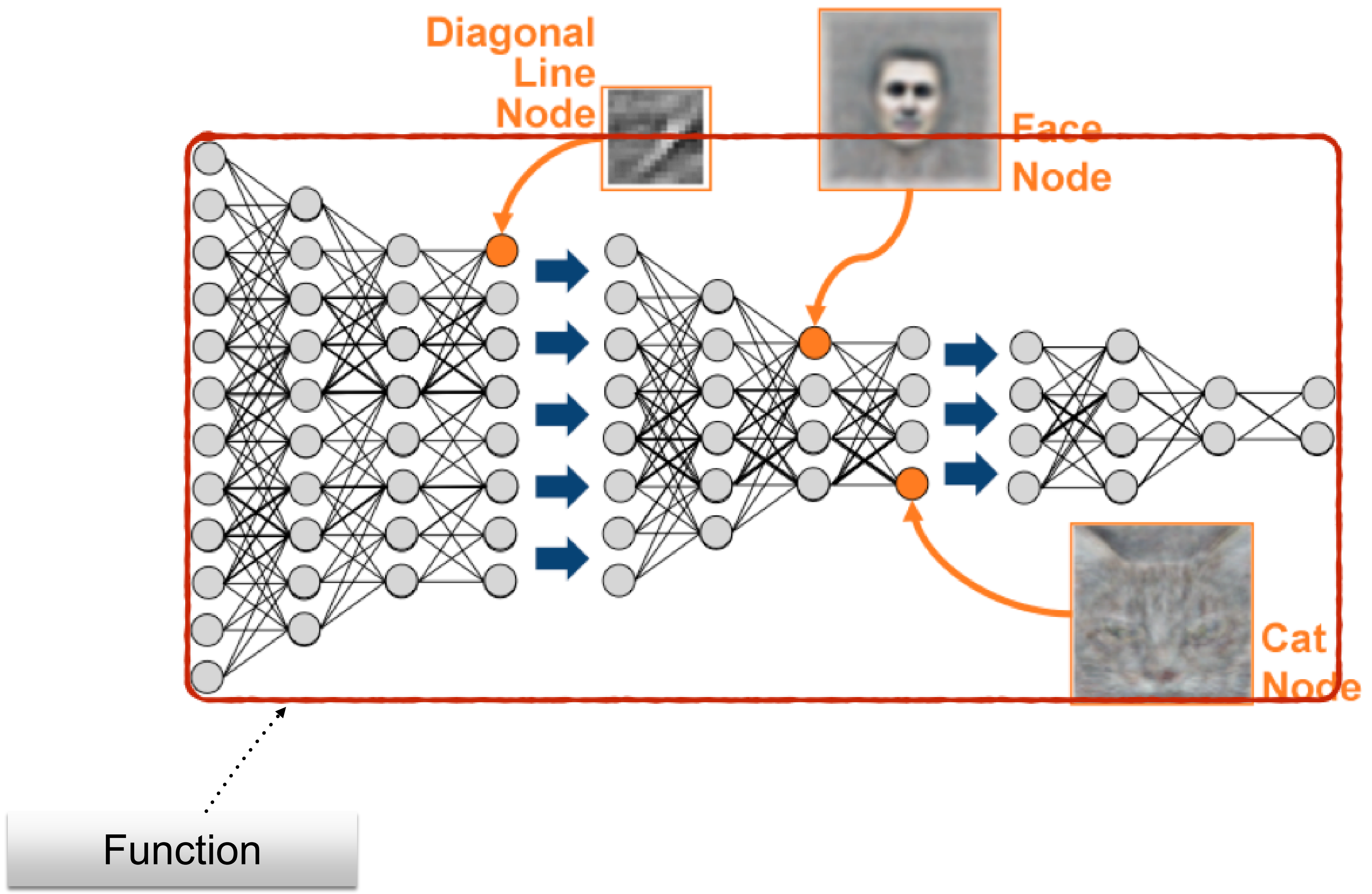


Problem 2: Inference over millions of dimensions? 😱

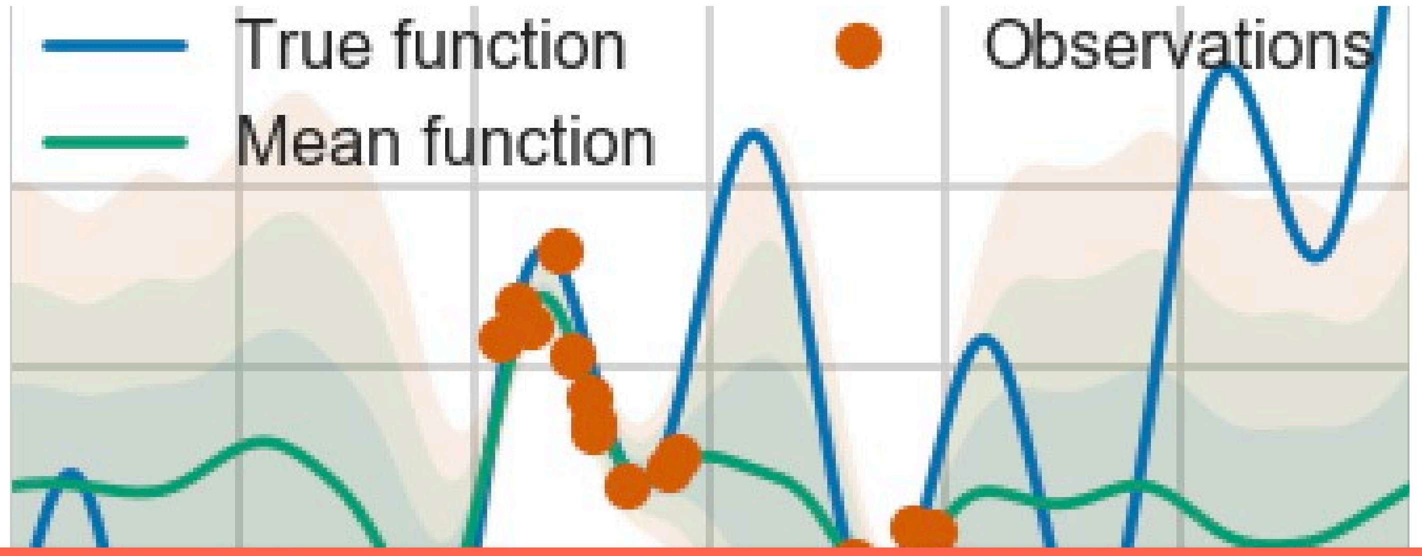


Stochastic Processes

- Can we bypass these limitations?
- Posit a **prior over functions** directly, rather than going through weights: stochastic process.
- **Gaussian Processes** is a prime example of stochastic processes



Gaussian Processes (GPs)



- GPs posit **priors over functions**: similar datapoints have similar predictions

Problem 1: (vanilla) GPs are not as flexible as neural nets for high dimensional tasks ☹□

- riors

Problem 2: GPs scale, in general, cubically w.r.t. the size of the dataset ☹□

- **Inference** is exact 😀

Combining the best of both worlds

- Can we **parametrize stochastic processes** that bypass the limitations of GPs?
- Yes! We only need to satisfy two necessary conditions (the **Kolmogorov extention theorem**)

1. Exchangeability: **p(**  ,  ,...,  **)=p(**  ,  ,...,  **)**

2. Consistency: **p(**  ,  ,...,  **)=∫p(**  ,  ,...,  , ) d 

De Finetti's Theorem

(cool maths)

Theorem 2 (De Finetti, 1930s). *A sequence of random variables* $(x_1, x_2, \ldots)$ *is infinitely exchangeable iff, for all* n,

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta),$$

(Source: Tamara Broderick)

for some measure P *on* θ.

Non-parametric Bayesian modeling

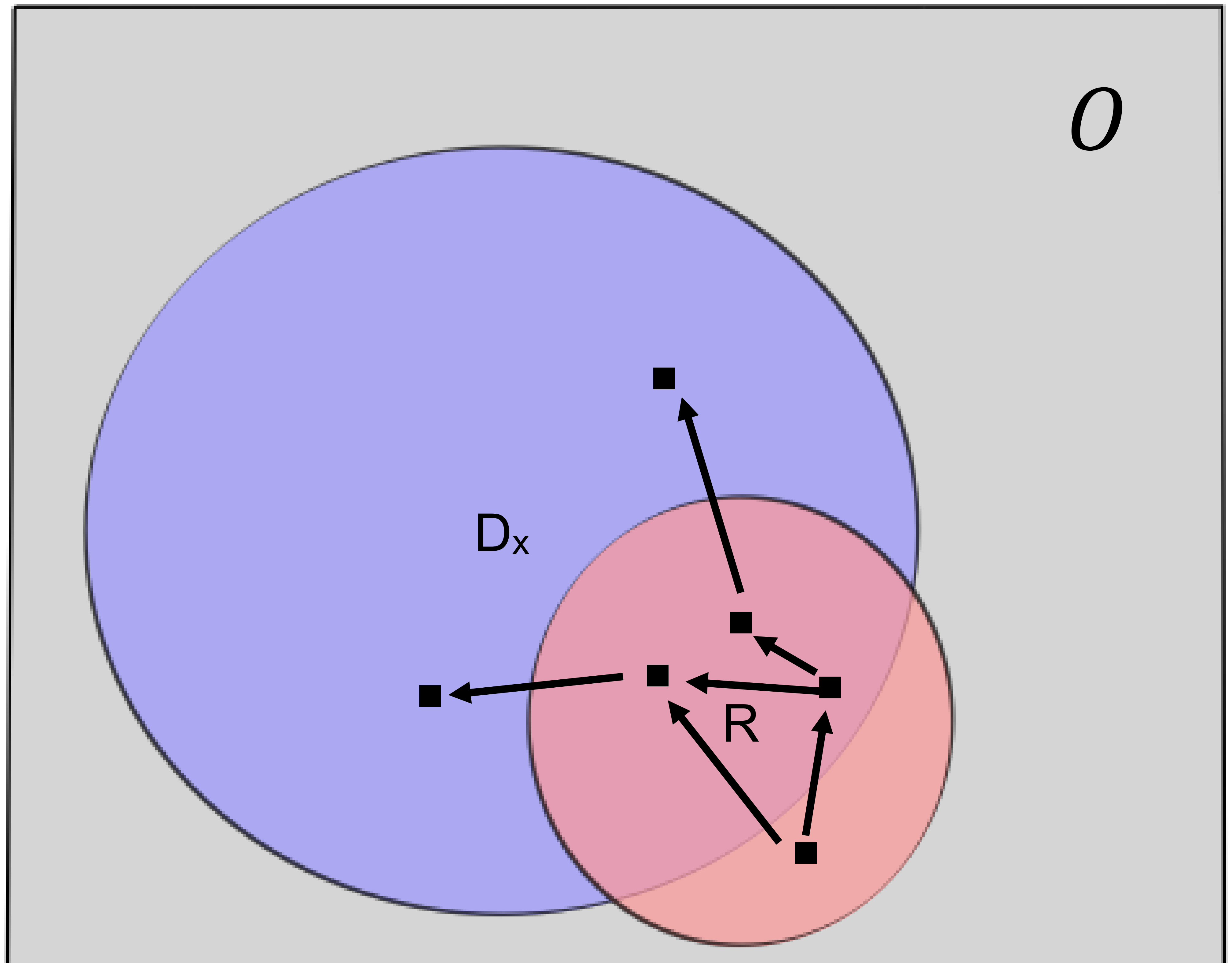
Bayesian modeling

Idea: Model Relational Structure

- **Exchangeable** joint model over all data cases
- Organize data in a **directed acyclic graph**
- Predictions depend on **parents** in this graph

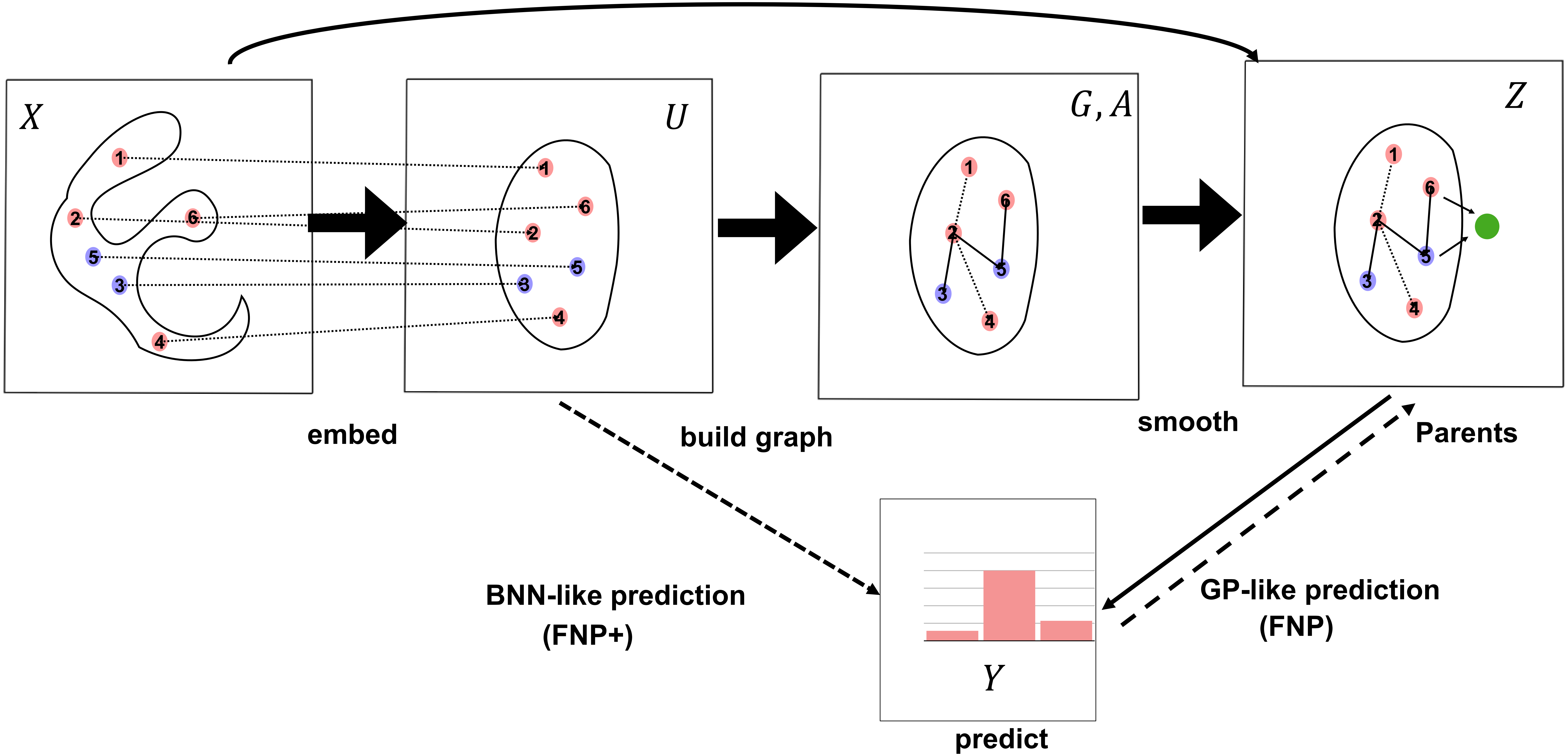
Building the Model

- Let $D = \{(x_i, y_i)\}_{i=1:N}$ be our training dataset and $D_x = \{x_i\}_{i=1:N}$ the training inputs
- Adopt a 'reference' set of input points $R = \{r_1, \ldots, r_n\}$ (similar to the 'inducing inputs' frequently used in GPs)
- We infer a DAG over R
- Data in $D_x \setminus R$ are leaf nodes of DAG



Venn diagram. R is not necessarily a subset of D_x, but for simplicity we will assume that it is.

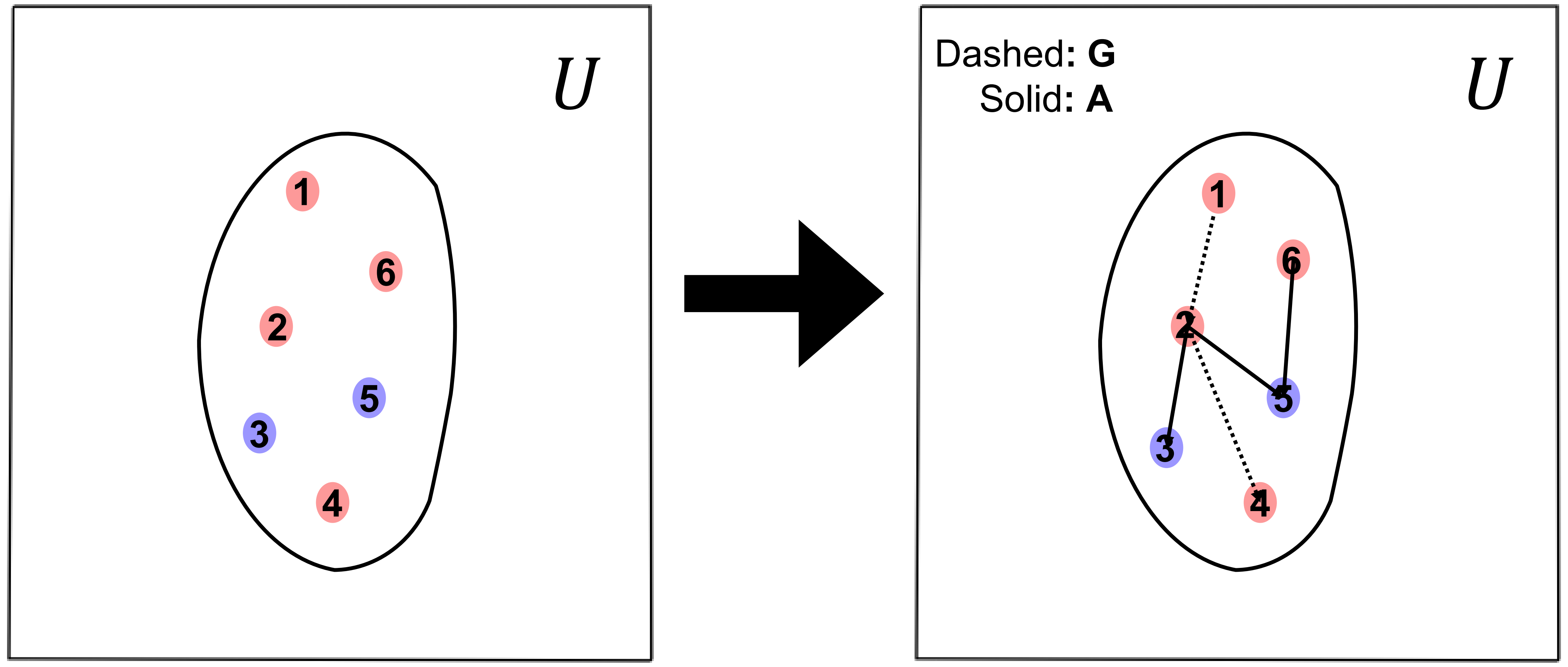
Building the Model: Overview



X
U
G, A
Z
Y
FNP
Xr
Xm
Ur
Par(G)
Um
G
A
Par(A)
Zr
Zm
Par(G)
Par(A)
Yr
Ym
NN embedding
"Graphon constructions"
Exchangeable functions of parents
Simple predictor
FNP+
Reference points
Other trn/tst points
Reference points
Other trn/tst points

Generalized Graphons

$p(\mathbf{G}, \mathbf{A} | \mathbf{U})$



Constructing a graph of dependencies among the points in U space

$$p(\mathbf{A}|\mathbf{U}) = \prod_{i \in D_x \setminus R} \prod_{j \in R} Bern(g(\mathbf{u}_i, \mathbf{u}_j))$$

$$p(\mathbf{G}|\mathbf{U}) = \prod_{i \in R} \prod_{j \in R, j \neq i} Bern(\mathbb{I}[t(\mathbf{u}_i) > t(\mathbf{u}_j)] g(\mathbf{u}_i, \mathbf{u}_j))$$

Topological ordering

$$t(\mathbf{u}_i) = \sum_k \log CDFNormal(\mathbf{u}_{ik})$$

Optimizing the model: Variational inference

Maximize the following ELBO to the marginal likelihood of D w.r.t. θ and variational parameters φ

In general, no (unbiased) minibatching ☹

$\mathcal{L}_R + \mathcal{L}$

Allows for unbiased minibatching! 😬

$$\mathcal{L}_R = \mathbb{E}_{p_\theta(\mathbf{U}_R,\mathbf{G}|\mathbf{X}_R)q_\phi(\mathbf{Z}_R|\mathbf{X}_R)}[\log p_\theta(\mathbf{y}_R,\mathbf{Z}_R|R,\mathbf{G}) - \log q_\phi(\mathbf{Z}_R|\mathbf{X}_R)]$$

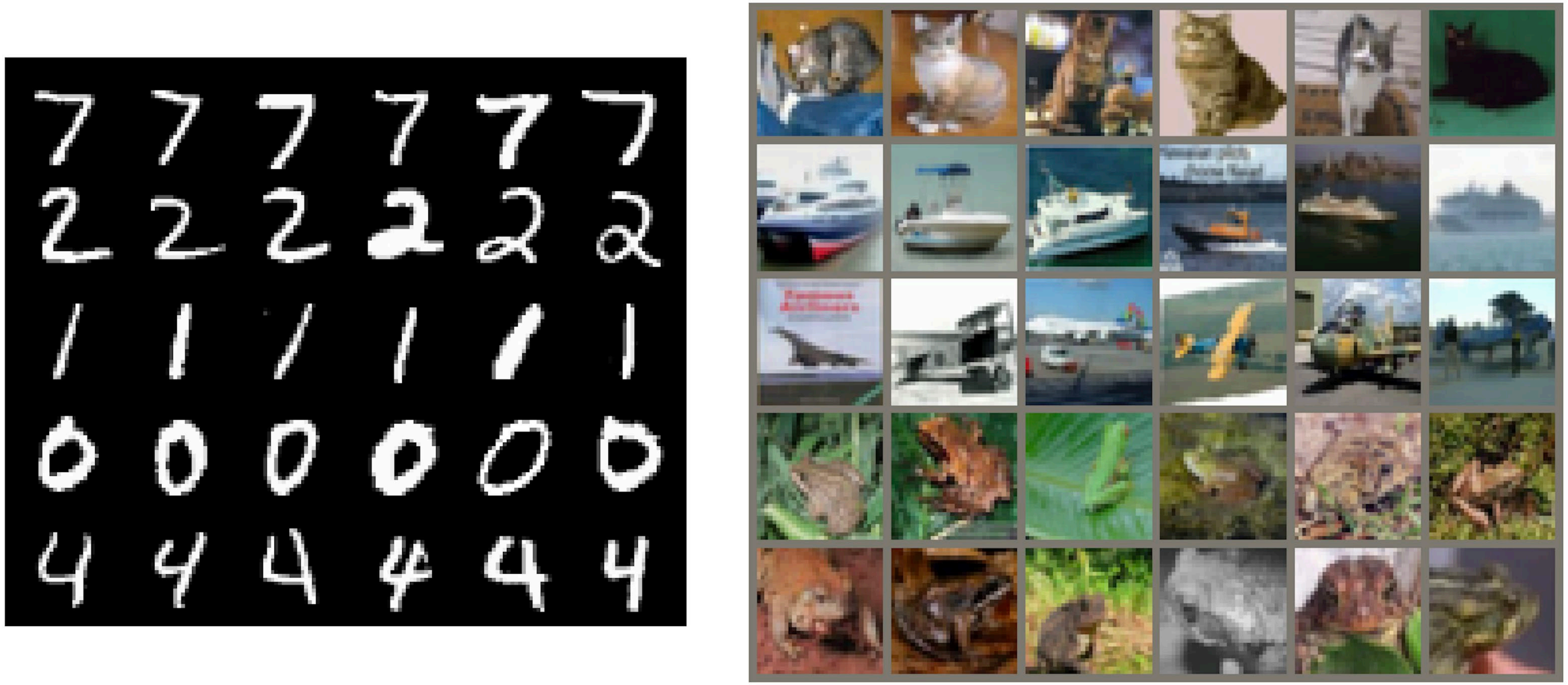
$$\mathcal{L}_{D_x \setminus R} = E_{p_\theta(\mathbf{U}_D,\mathbf{A}|\mathbf{X}_D)q_\phi(\mathbf{Z}_M|\mathbf{X}_M)}[\log p_\theta(\mathbf{y}_M|\mathbf{Z}_M) + \log p_\theta(\mathbf{Z}_M|par_{\mathbf{A}}(R,\mathbf{y}_R)) - \log q_\phi(\mathbf{Z}_M|\mathbf{X}_M)]$$

- Where we assumed that:

$$q_\phi(\mathbf{U}_D,\mathbf{G},\mathbf{A},\mathbf{Z}_D|\mathbf{X}_D,\mathbf{y}_D) = p_\theta(\mathbf{U}_D|\mathbf{X}_D)p(\mathbf{G}|\mathbf{U}_R)p(\mathbf{A}|\mathbf{U}_D)q_\phi(\mathbf{Z}_D|\mathbf{X}_D)$$

$$q_\phi(\mathbf{Z}_D|\mathbf{X}_D) = \prod_i q_\phi(\mathbf{z}_i|\mathbf{x}_i)$$

Example graphs of dependencies

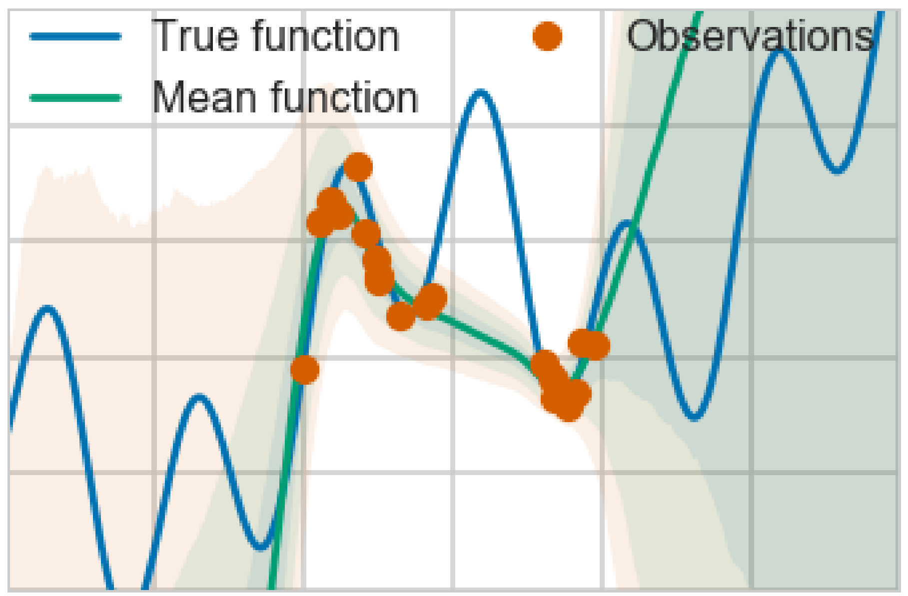


Examples of the bipartite graph **A** that the FNP learns. The first column of each image is a query point and the rest are the five most probable parents from the reference set R. We can see that the FNP associates same class inputs.

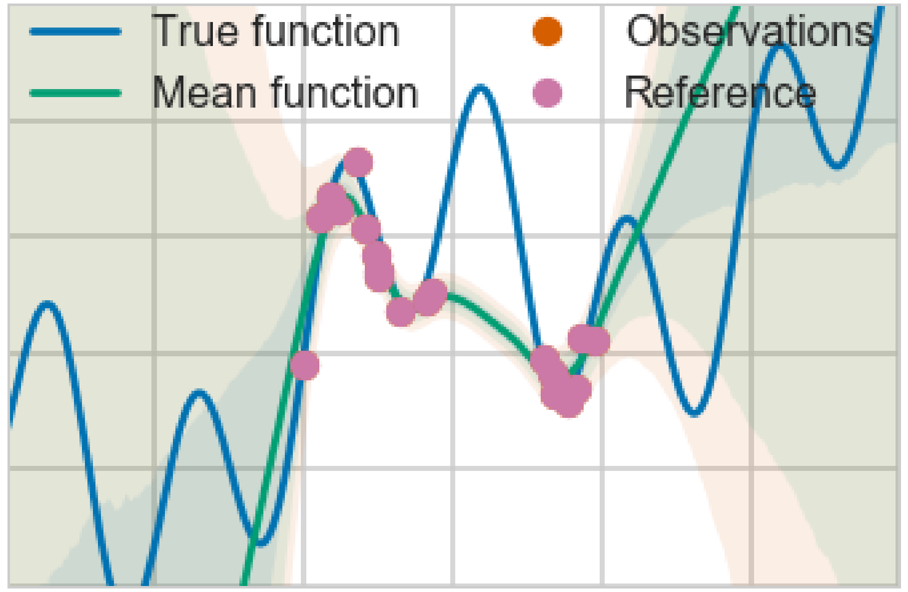
Example graphs of dependencies

A DAG over R on MNIST, obtained after propagating the means of **U** and thresholding edges that have less than 0.5 probability in **G**. We can see that FNP learns a meaningful **G** by connecting points that have the same class.

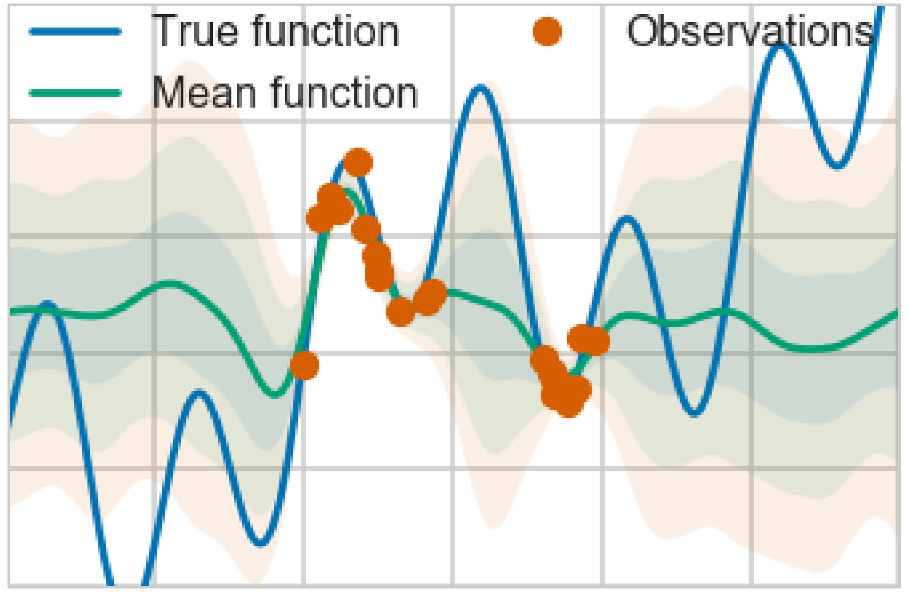
Inductive biases in toy regression



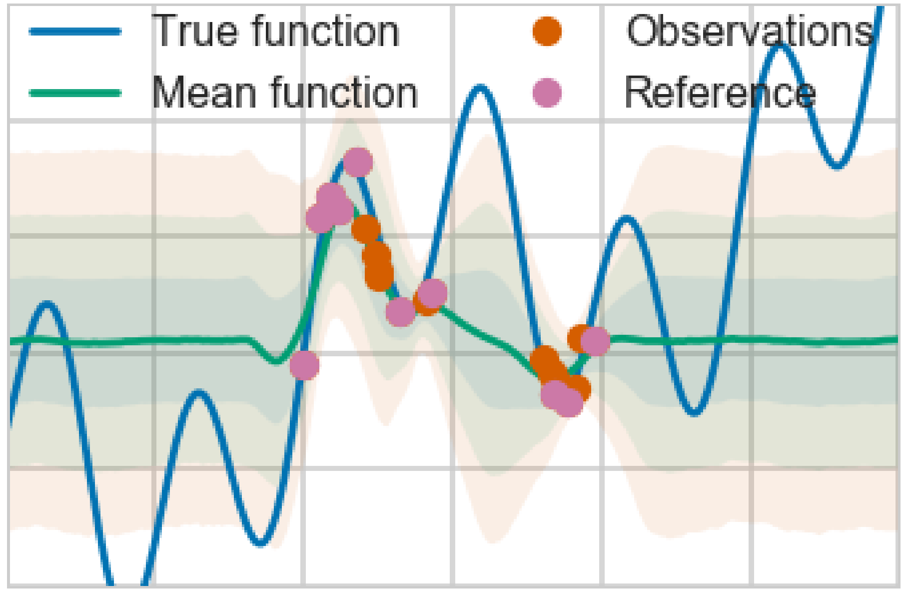
MC-Dropout



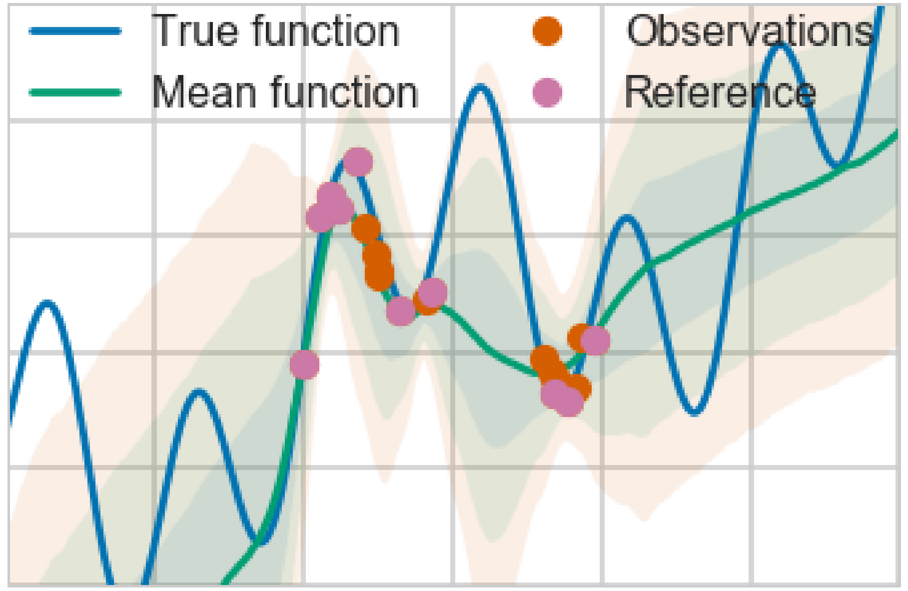
Neural Process



Gaussian Process



FNP



FNP+

Literature FNP

[1] *Neural Processes*, Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola Danilo J. Rezende S. M. Ali Eslami, Yee Whye Teh, https://arxiv.org/pdf/1807.01622.pdf
[2] *Auto-encoding Variational Bayes*, Diederik P. Kingma, Max Welling, https://arxiv.org/pdf/1312.6114.pdf
[3] *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*, Danilo J. Rezende, Shakir Mohamed, Daan Wierstra, https://arxiv.org/pdf/1401.4082.pdf
[4] Deep Variational Information Bottleneck, Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, Kevin Murphy, https://arxiv.org/pdf/1612.00410.pdf
[5] Associative Compression Networks for Representation Learning, Alex Graves, Jacob Menick, Aaron van den Oord, https://arxiv.org/pdf/1804.02476.pdf
[6] Few-shot Generative Modelling with Generative Matching Networks, Sergey Bartunov, Dmitry P. Vetrov, http://proceedings.mlr.press/v84/bartunov18a/bartunov18a.pdf
[7] Matching Networks for One-shot Learning, Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra, https://arxiv.org/pdf/1606.04080.pdf

Punchline

Functional Priors are more intuitive and can still be scalable for both supervised and unsupervised learning.