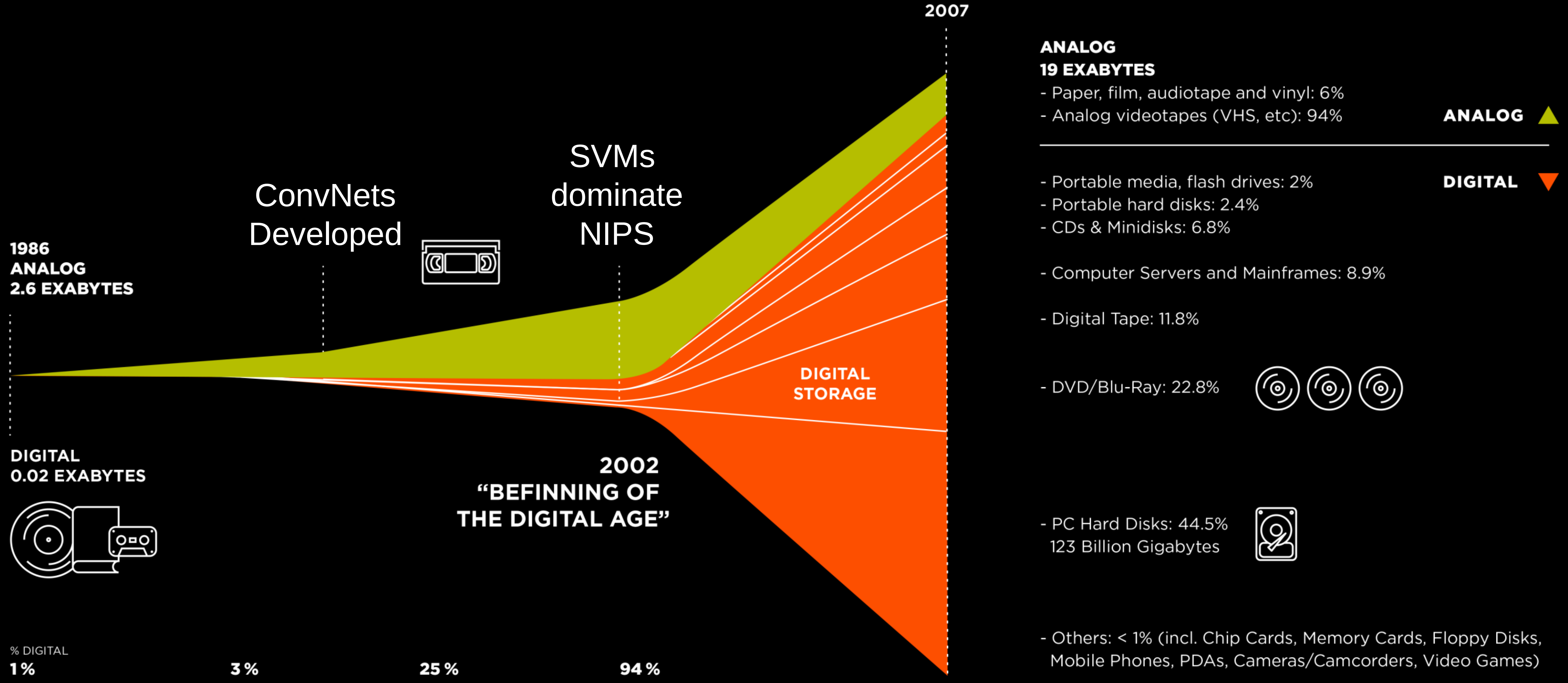
Uncertainty Propagation

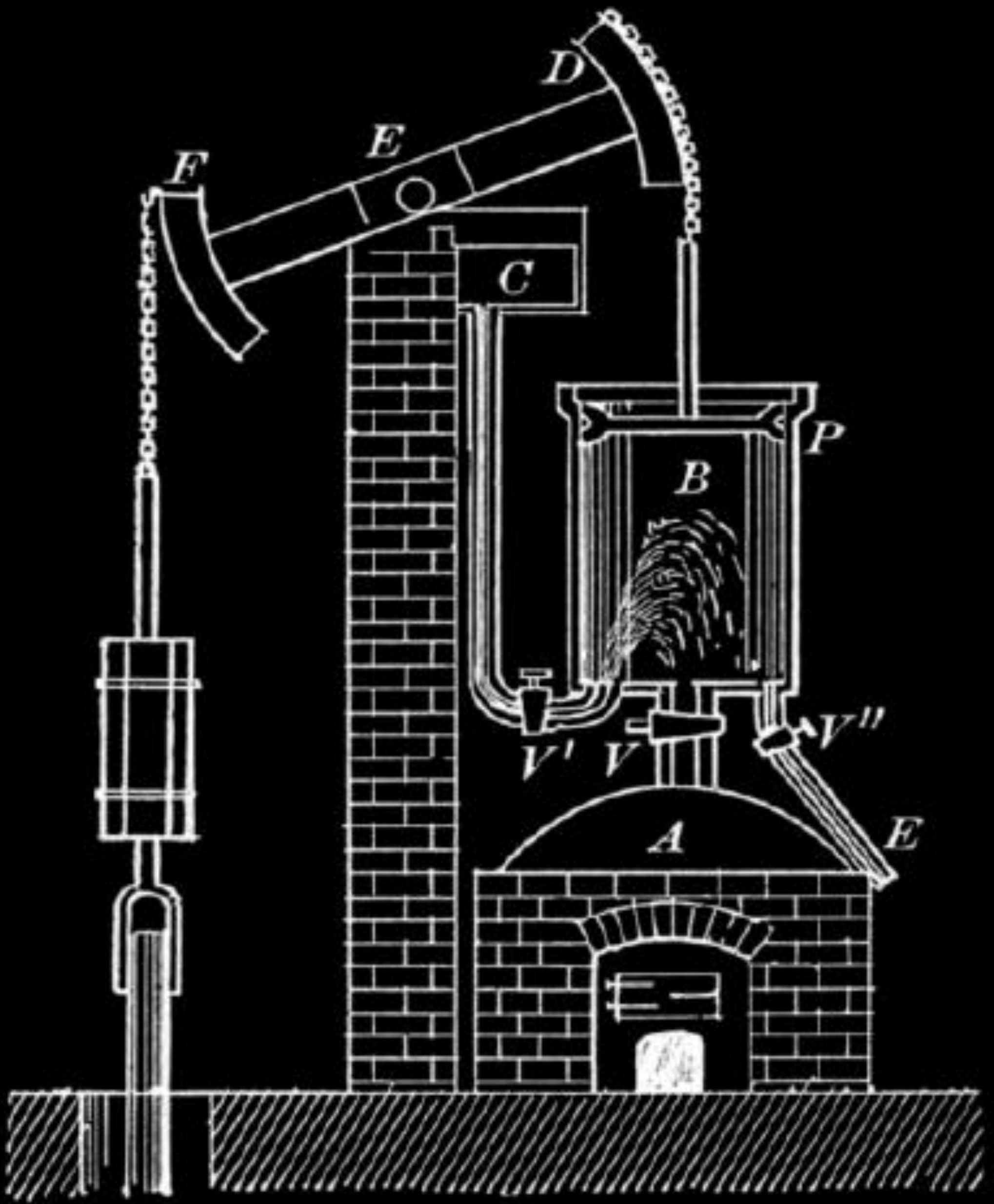
NEIL LAWRENCE
UNIVERSITY OF SHEFFIELD
@lawrennd

GLOBAL INFORMATION STORAGE CAPACITY
IN OPTIMALLY COMPRESSED BYTES

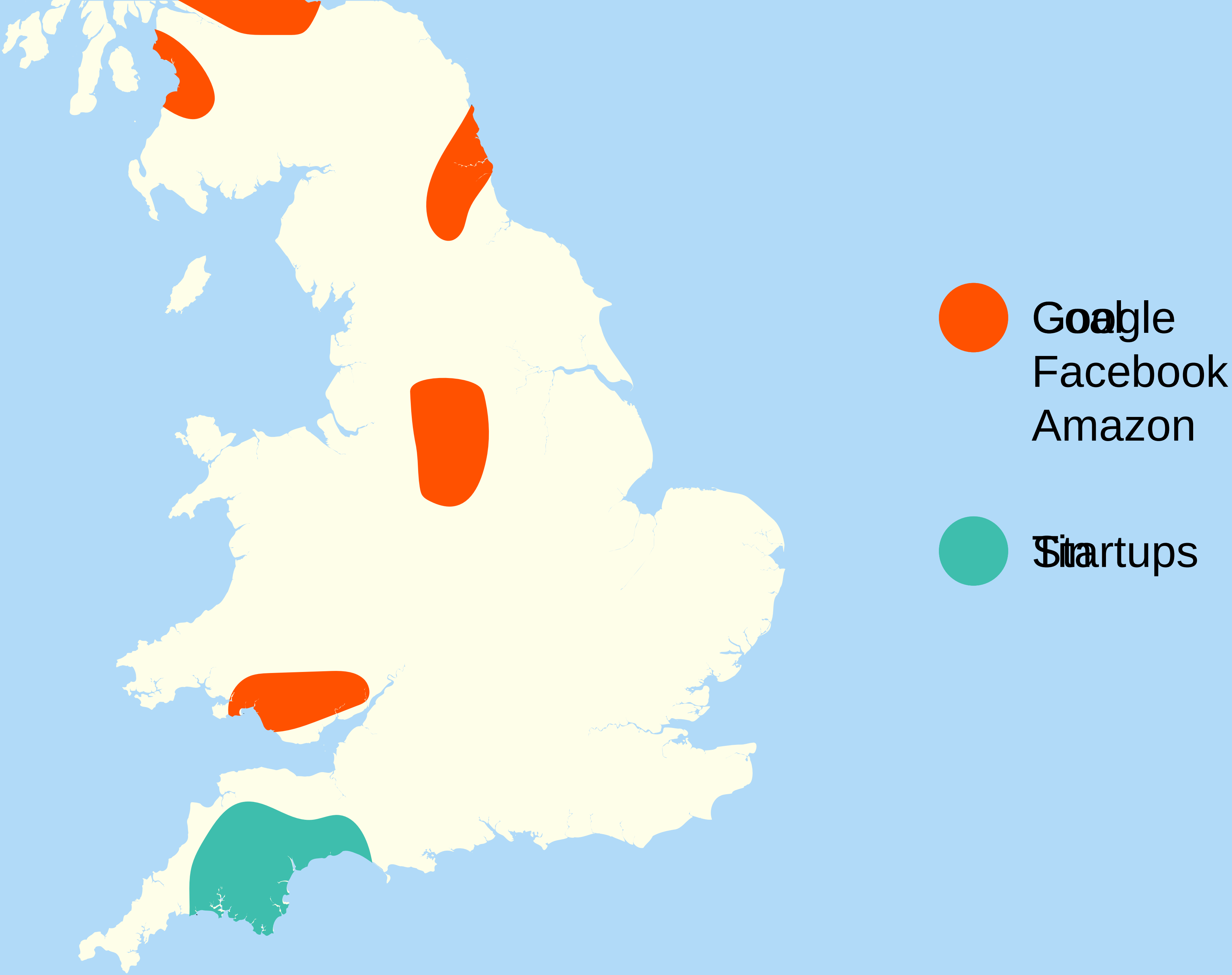


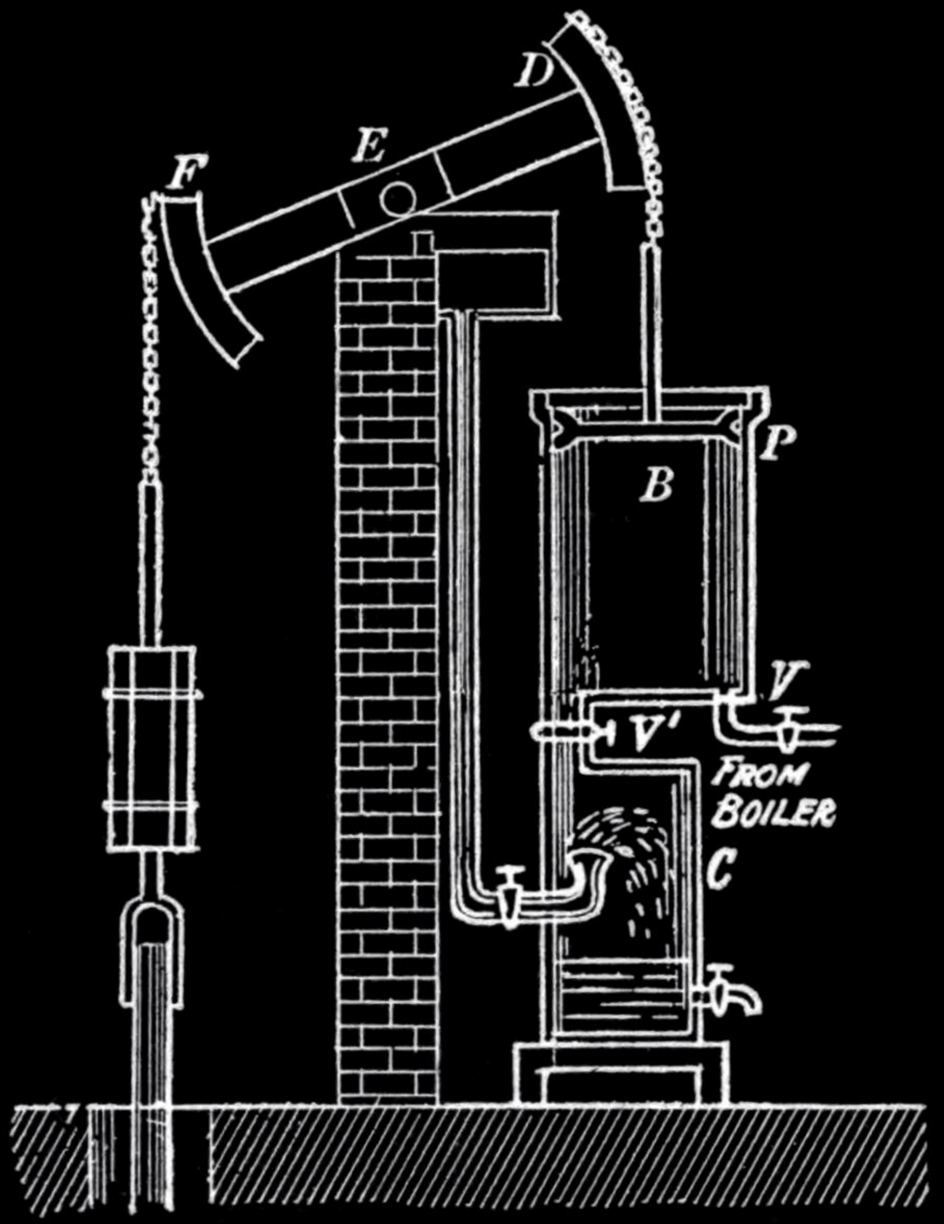
Source: Hilbert, M., & López, P. (2011). The World’s Technological Capacity to Store, Communicate, andCompute Information. Science, 332 (6025), 60-65. martinhilbert.net/worldinfocapacity.html

DIGITAL
280 EXABYTES

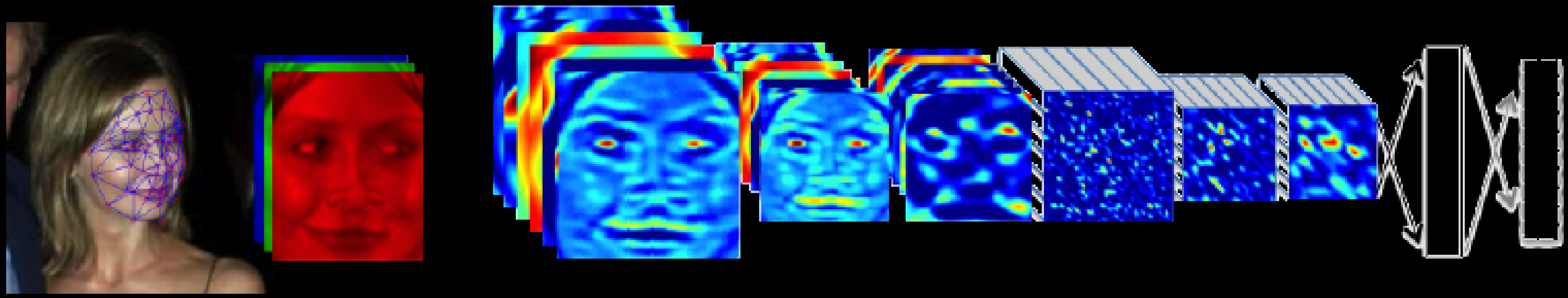


D
E
F
C
P
B
V'
V
V''
A
E


Google
Facebook
Amazon
Startups

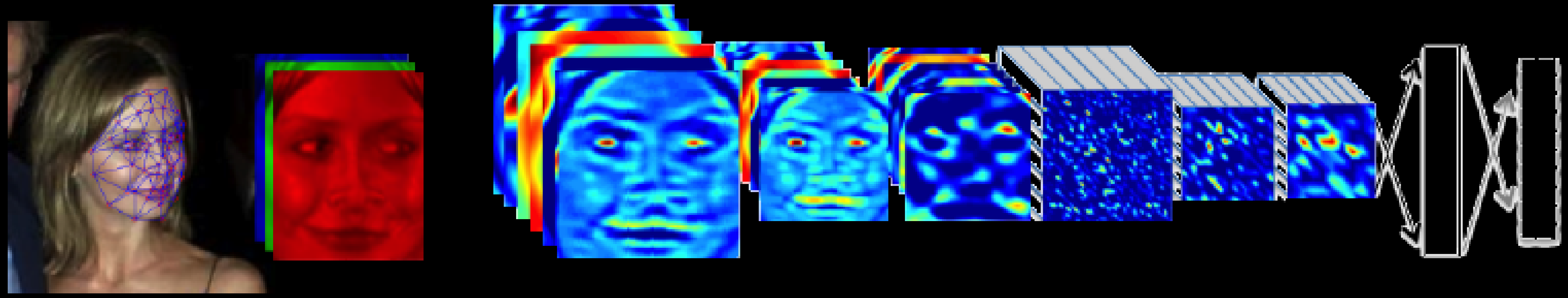

D
E
F
P
B
V
V'
FROM BOILER
C

Outline of the DeepFace architecture. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Color illustrates feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.



Source: DeepFace

$$\mathbf{g}(x)$$



$\mathbf{f}_1(x)$ $\mathbf{f}_2(\cdot)$ $\mathbf{f}_3(\cdot)$ $\mathbf{f}_4(\cdot)$ $\mathbf{f}_5(\cdot)$ $\mathbf{f}_6(\cdot)$ $\mathbf{f}_7(\cdot)$ $\mathbf{f}_8(\cdot)$ $\mathbf{f}_9(\cdot)$

$$\mathbf{g}(x) = \mathbf{f}_9\left(\mathbf{f}_8\left(\mathbf{f}_7\left(\mathbf{f}_6(\cdots)\right)\right)\right)$$

$$\mathbf{f}_9(\mathbf{h}) = \begin{bmatrix} \phi(\sum_i w_{1i} h_i) \\ \phi(\sum_i w_{2i} h_i) \\ \vdots \\ \phi(\sum_i w_{ki} h_i) \end{bmatrix}$$

$$\mathbf{f}_9(\mathbf{h}) = \phi(\mathbf{W}\mathbf{h})$$

$$\mathbf{W} \in \Re^{k_8 \times k_9}$$

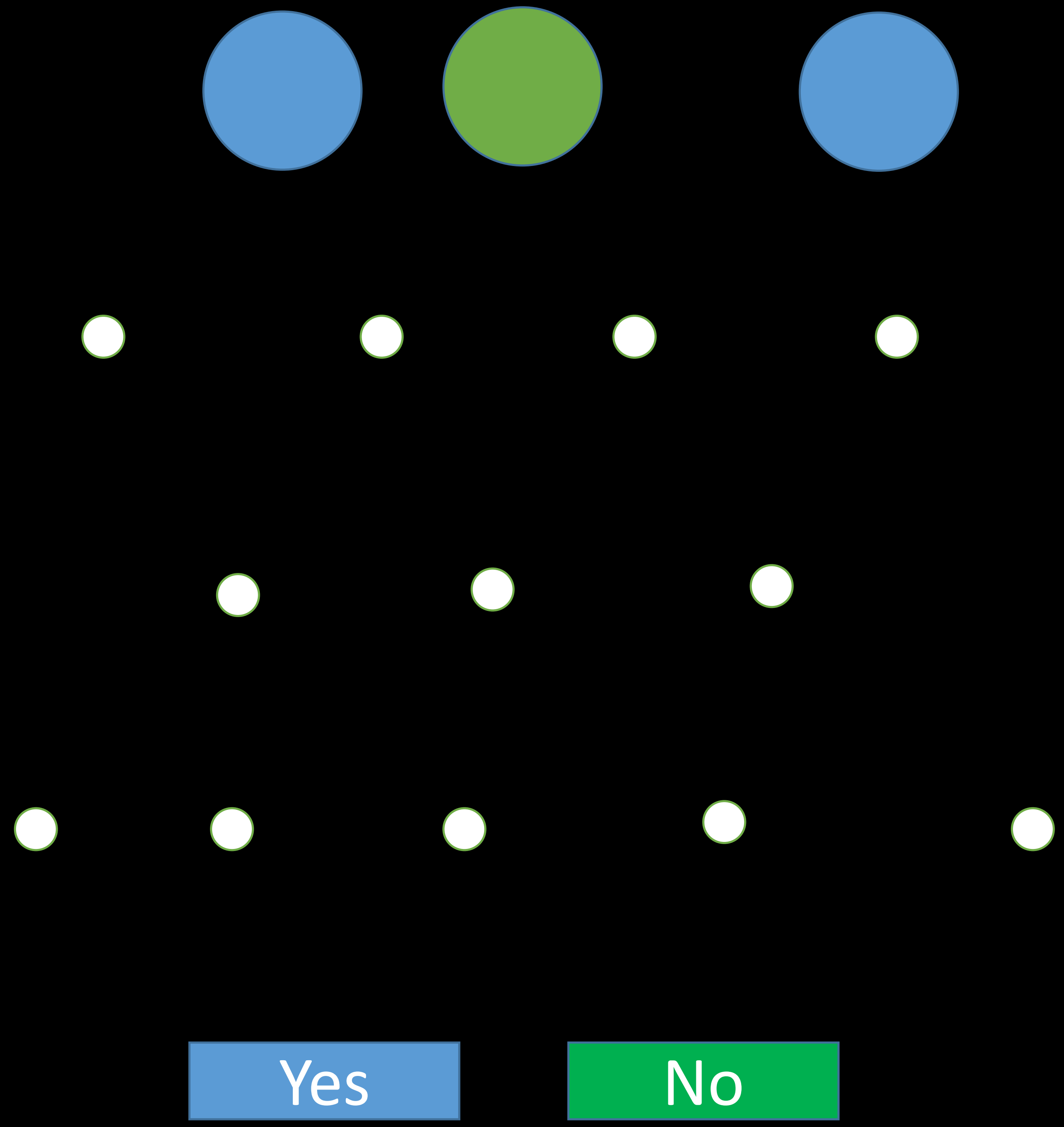
$\mathbf{x}$

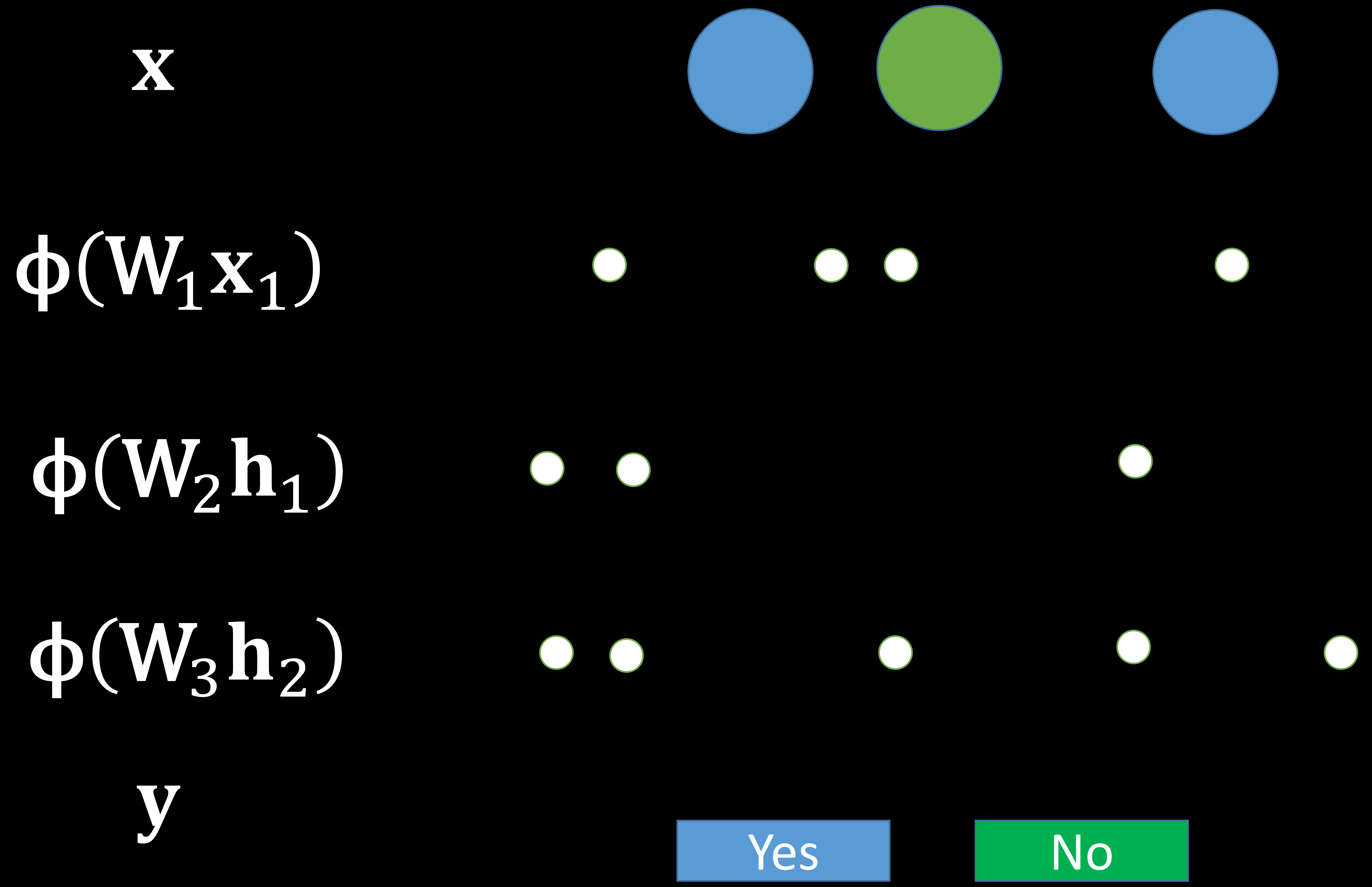
$\phi(\mathbf{W}_1 \mathbf{x}_1)$

$\phi(\mathbf{W}_2 \mathbf{h}_1)$

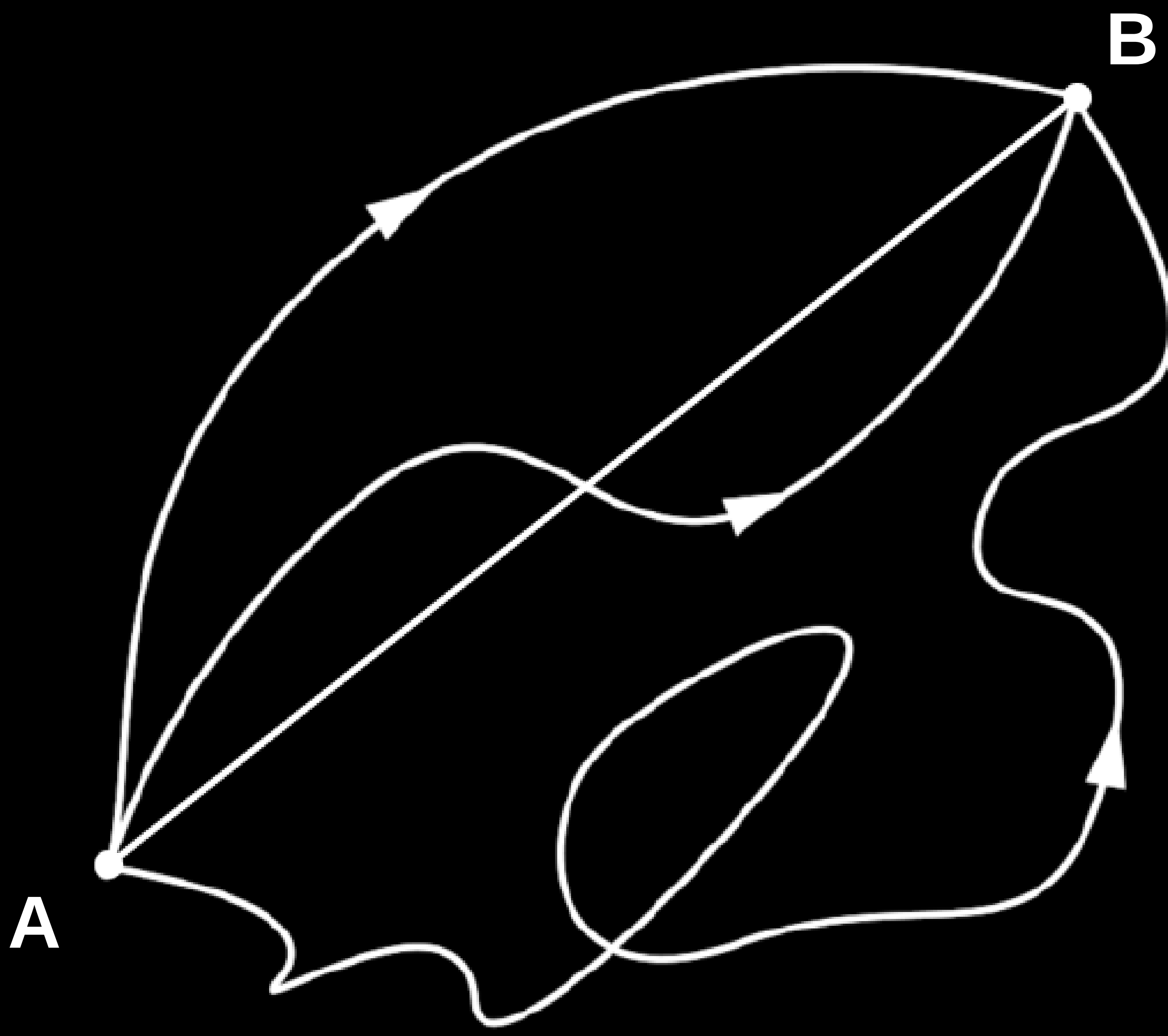
$\phi(\mathbf{W}_3 \mathbf{h}_2)$

$\mathbf{y}$

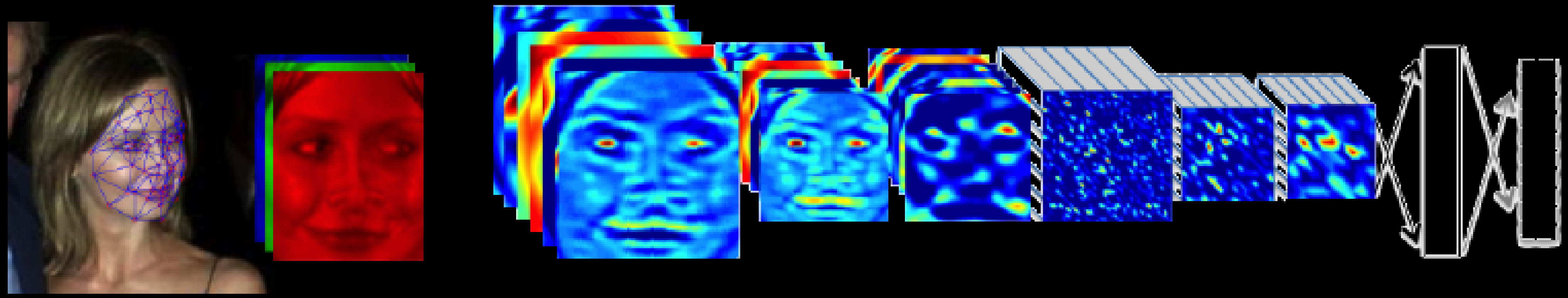




$\mathbf{x}$
$\phi(\mathbf{W}_1\mathbf{x}_1)$
$\phi(\mathbf{W}_2\mathbf{h}_1)$
$\phi(\mathbf{W}_3\mathbf{h}_2)$
$\mathbf{y}$
Yes
No


B
A

$$g(x)$$



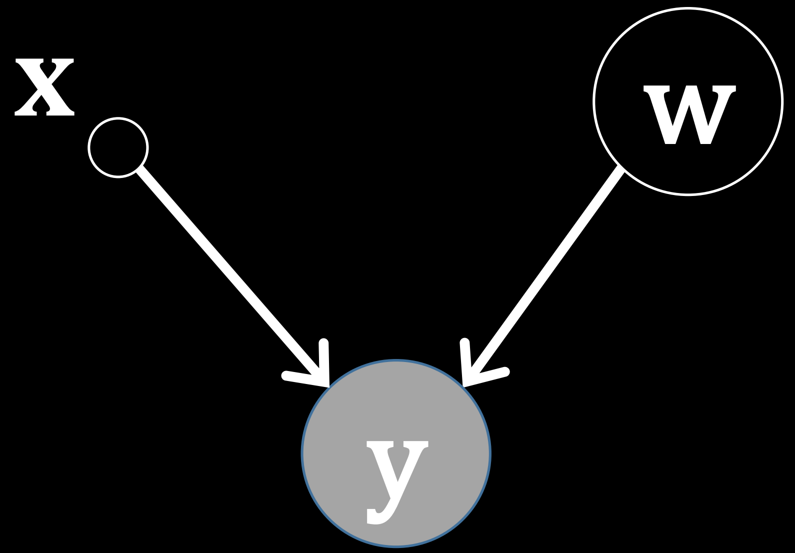
$$\frac{\mathrm{d}g(x)}{\mathrm{d}x}$$

$$\int g(x)p(x)\mathrm{d}x$$

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - g(\mathbf{x}_i; \mathbf{w})\right)^2$$

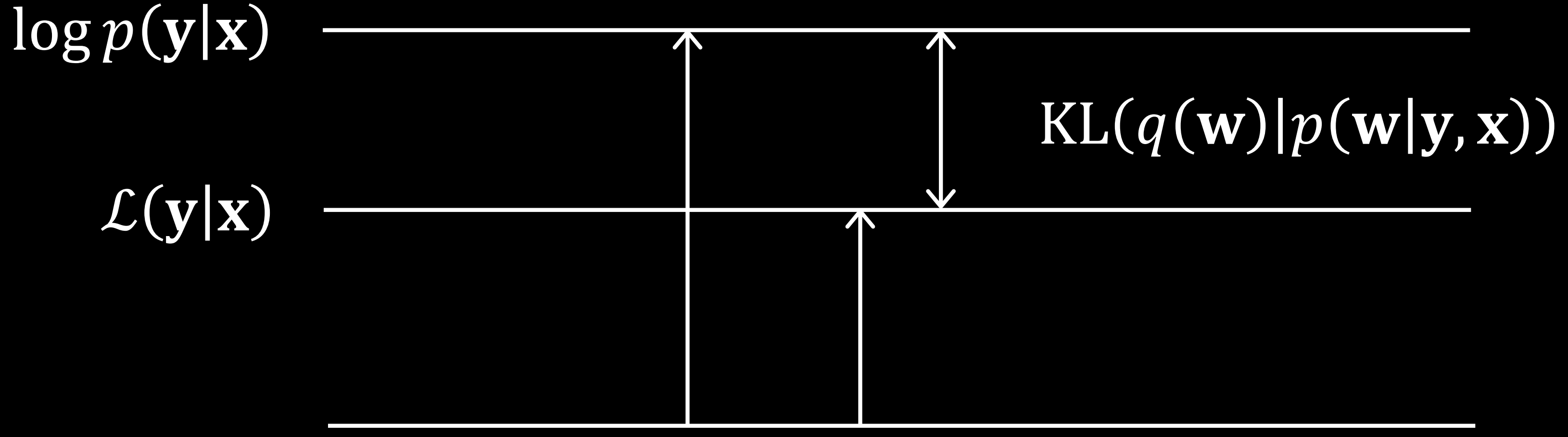
$$\log p(\mathbf{y}|\mathbf{w}, \mathbf{x}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - g(\mathbf{x}_i; \mathbf{w})\right)^2 + \frac{n}{2} \log 2\pi\sigma^2$$

$$p(\mathbf{y}, \mathbf{w}|\mathbf{x}) = p(\mathbf{y}|\mathbf{w}, \mathbf{x})p(\mathbf{w})$$



$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{w}, \mathbf{x})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

$$\log p(\mathbf{y}|\mathbf{x}) \geq \int q(\mathbf{w}) \log \frac{p(\mathbf{y}|\mathbf{w},\mathbf{x})p(\mathbf{w})}{q(\mathbf{w})} \mathrm{d}\mathbf{w}$$



expected
log likelihood

dissimilarity
between $q(\mathbf{w})$
and $p(\mathbf{w})$

$$\mathcal{L}(\mathbf{y}|\mathbf{x}) = \left\langle \log p(\mathbf{y}|\mathbf{w}, \mathbf{x}) \right\rangle_{q(\mathbf{w})} - \mathrm{KL}(q(\mathbf{w}) \| p(\mathbf{w}))$$

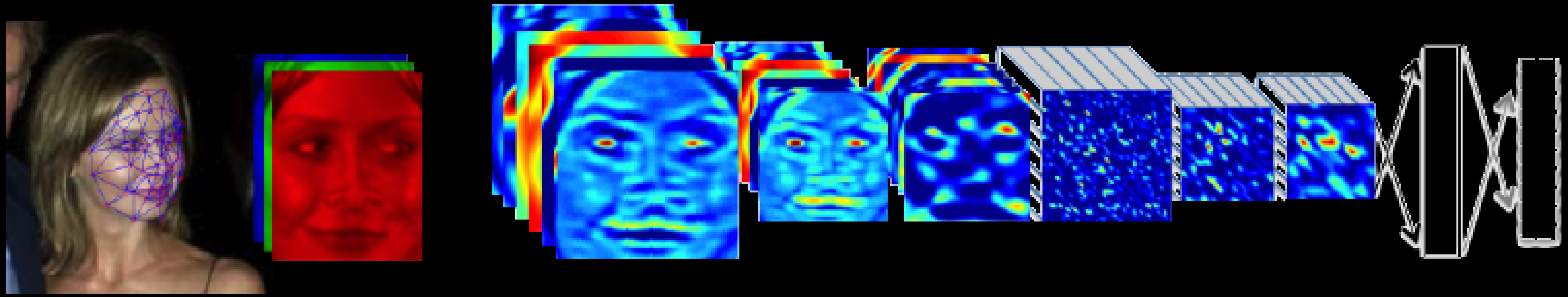
$$\mathcal{L}(\mathbf{y}|\mathbf{x}) = -\left\langle \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}_i; \mathbf{w}) \right)^2 \right\rangle_{q(\mathbf{w})} - \mathrm{KL}(q(\mathbf{w}) \| p(\mathbf{w})) + \mathrm{const}$$

Gaussian Processes

x_1, y_1 $\quad p(\mathbf{f}|\mathbf{x})$ $\quad x_2, y_2$

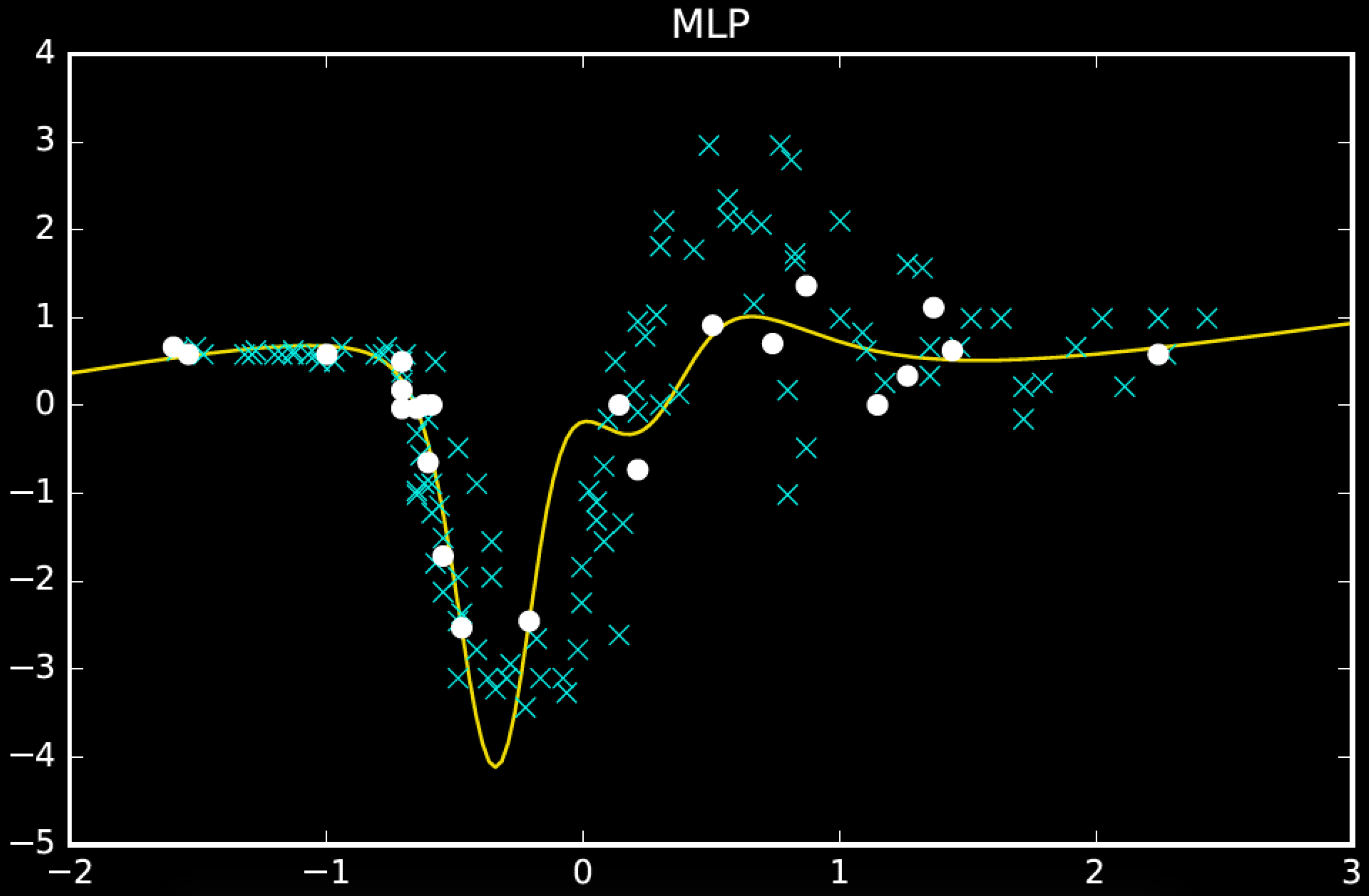
$p(y_1|f_1)$ $\quad p(\mathbf{f}|\mathbf{y}, \mathbf{x})$ $\quad p(y_2|f_2)$

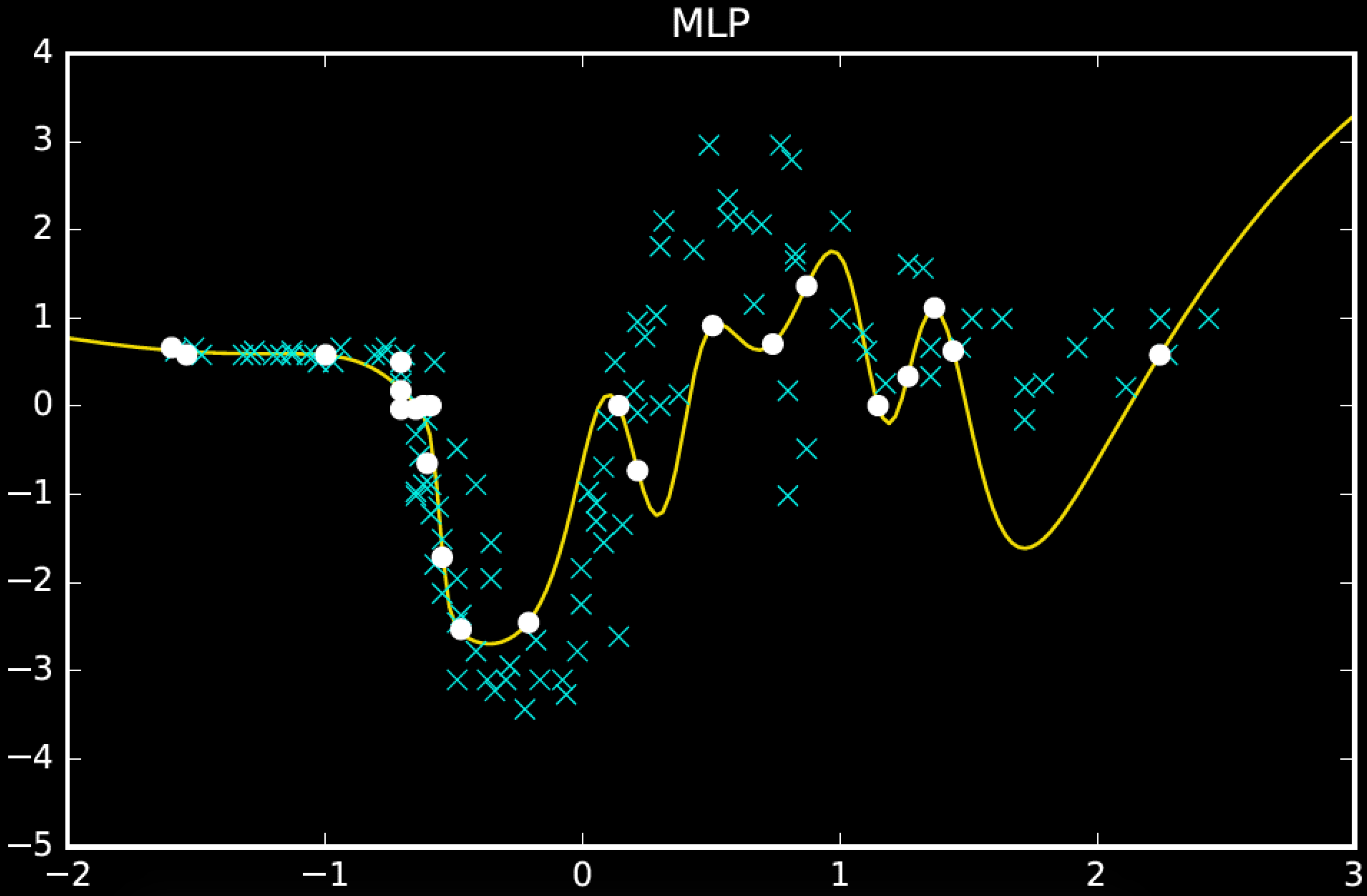
$$\mathbf{g}(x)$$

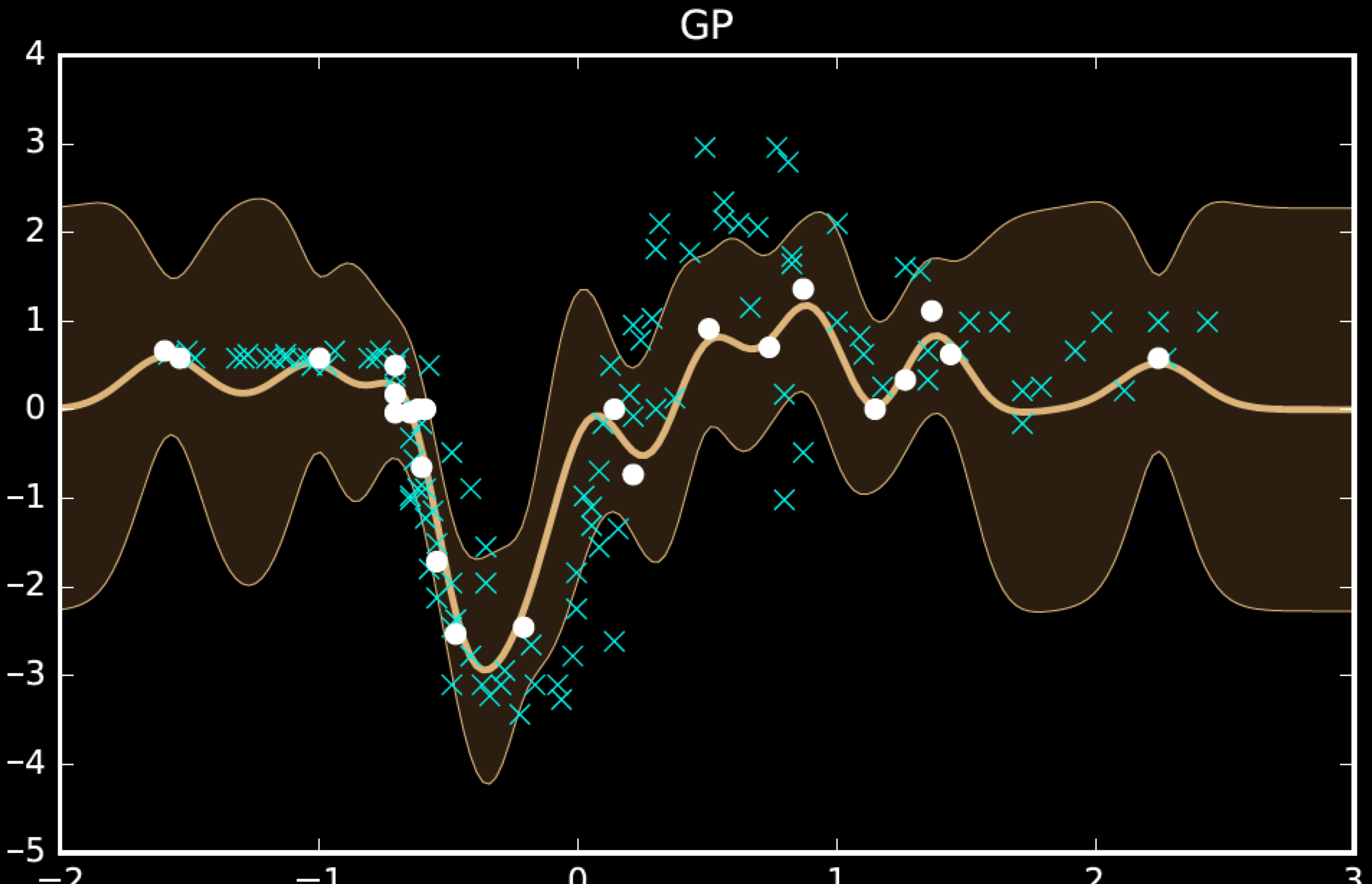


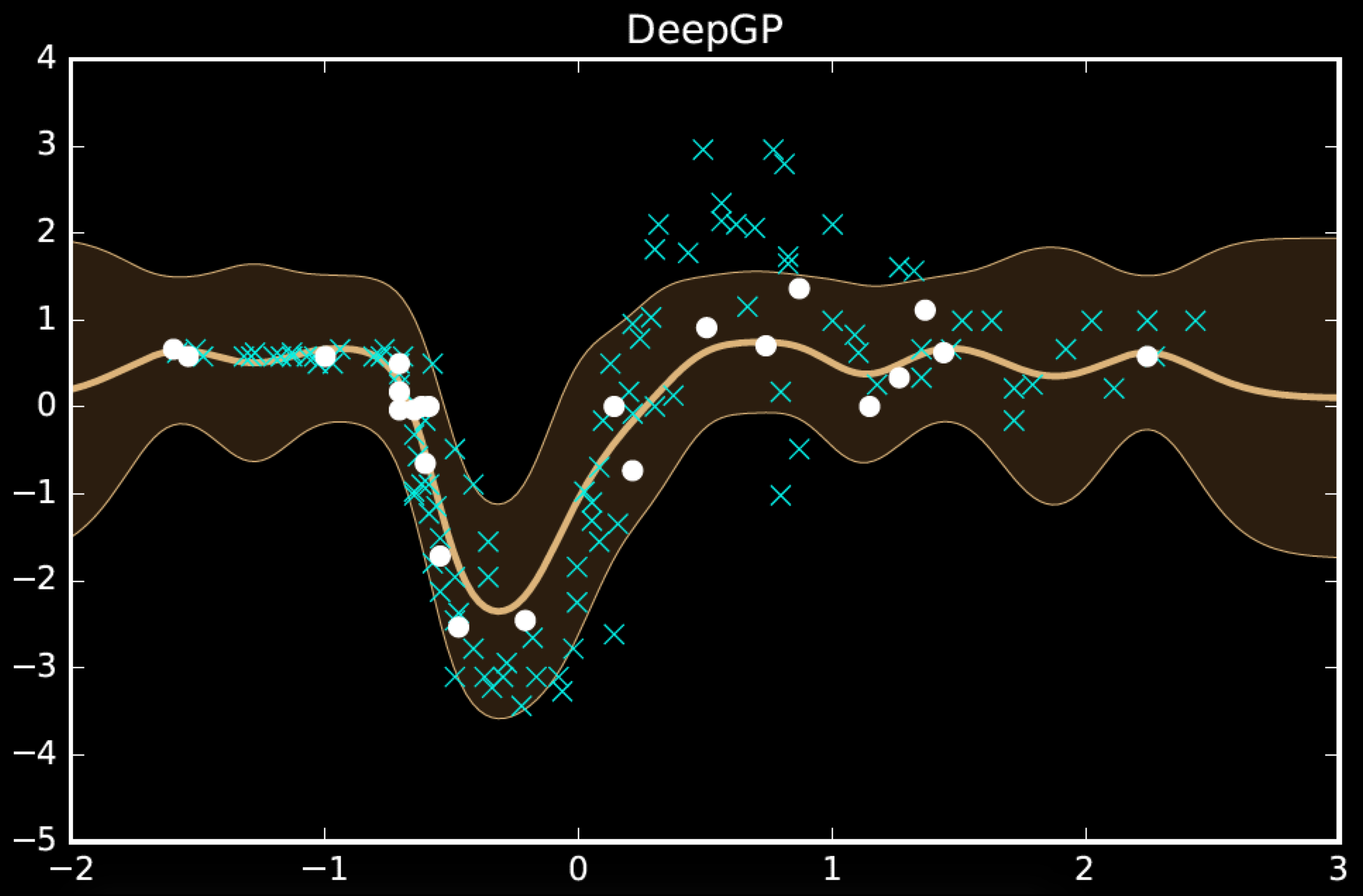
$\mathbf{f}_1(x)$ $\mathbf{f}_2(\cdot)$ $\mathbf{f}_3(\cdot)$ $\mathbf{f}_4(\cdot)$ $\mathbf{f}_5(\cdot)$ $\mathbf{f}_6(\cdot)$ $\mathbf{f}_7(\cdot)$ $\mathbf{f}_8(\cdot)$ $\mathbf{f}_9(\cdot)$

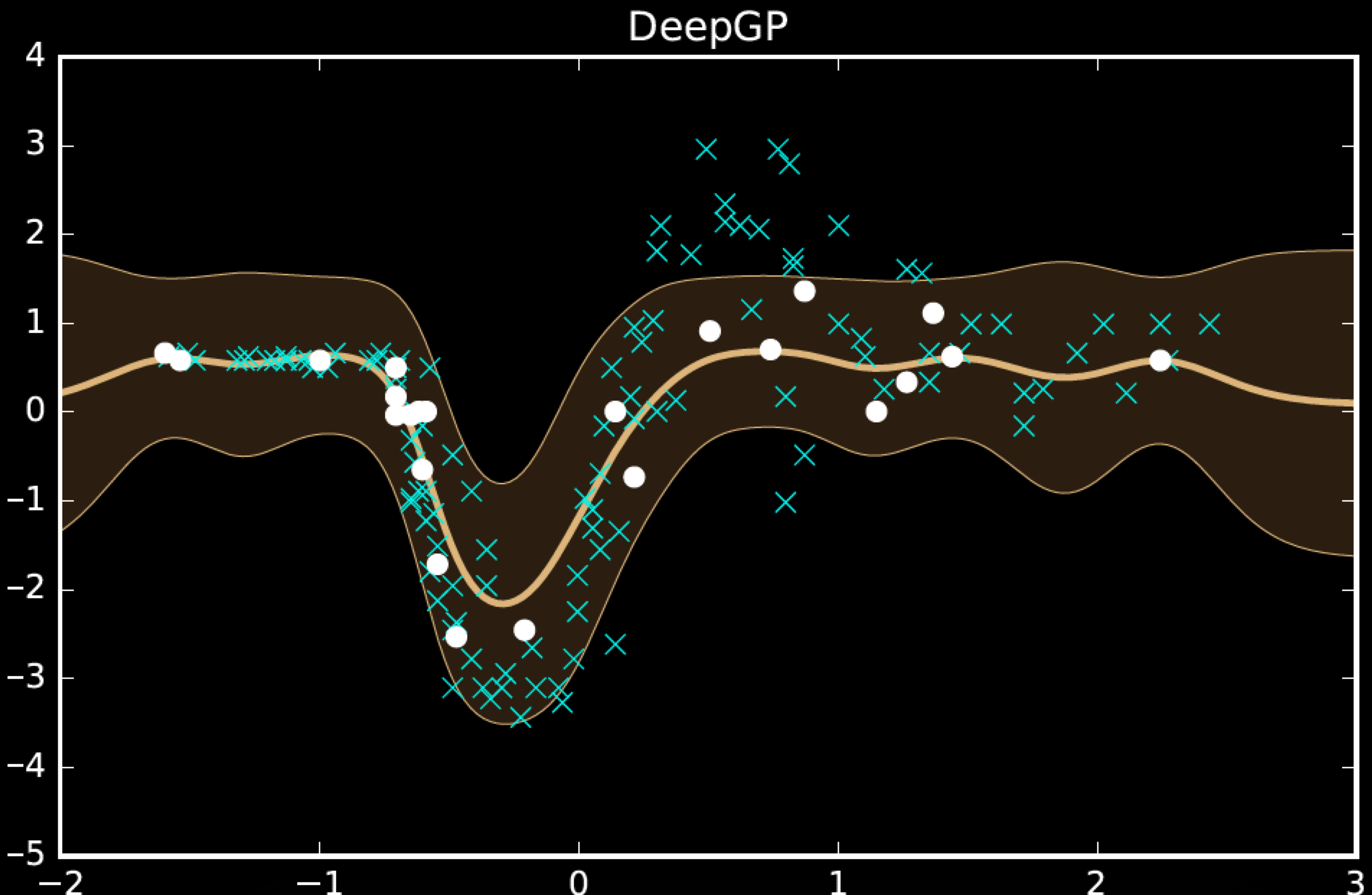
$$\mathbf{g}(x) = \mathbf{f}_9\left(\mathbf{f}_8\left(\mathbf{f}_7\left(\mathbf{f}_6(\cdots)\right)\right)\right)$$


MLP
4
3
2
1
0
−1
−2
−3
−4
−5
−2
−1
0
1
2
3


MLP
4
3
2
1
0
−1
−2
−3
−4
−5
−2
−1
0
1
2
3


GP
4
3
2
1
0
−1
−2
−3
−4
−5
−2
−1
0
1
2
3


DeepGP
4
3
2
1
0
−1
−2
−3
−4
−5
−2
−1
0
1
2
3



DeepGP
4
3
2
1
0
−1
−2
−3
−4
−5
−2
−1
0
1
2
3

model	MSE (train)	MSE (test)
mlp (200 iters)	108.5	1185.1
mlp (converged)	24.0	1338.2
gp	59.2	1095.4
deep gp (2)	146.2	833.7
deep gp (3)	182.5	843.6

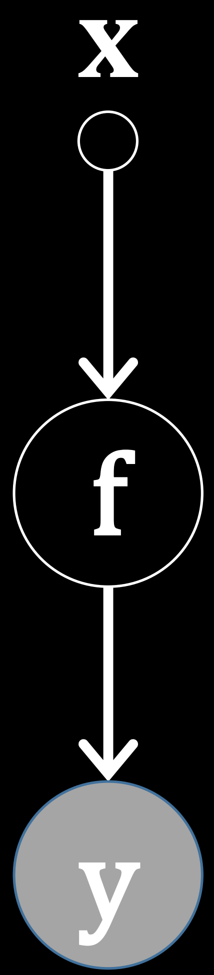
One hundred hidden nodes, one hundred inducing points

$$\mathbf{f}|\mathbf{x} \sim N\left(0, \mathbf{K}_{ff}\right)$$

$$k_{ff}(x_i, x_i') = \alpha \exp\left(-\frac{\left\|x_i - x_i'\right\|^2}{2\ell^2}\right)$$

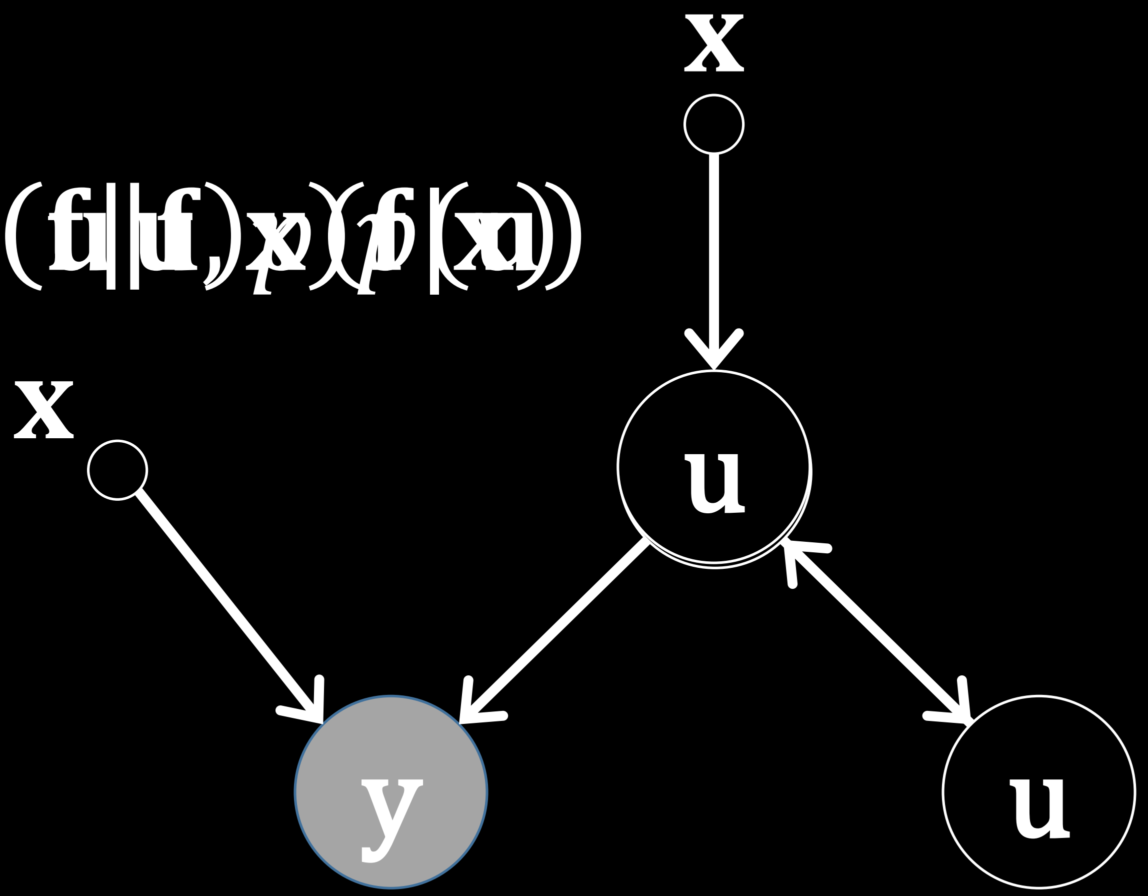
$$y_i|f_i \sim N(0, \sigma^2)$$

$$p(\mathbf{y}, \mathbf{f}|\mathbf{x}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x})$$



$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x})\mathrm{d}\mathbf{f}$$

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u}, \mathbf{x})p(\mathbf{u})$$



$$p(\mathbf{y}|\mathbf{u}, \mathbf{x})p(\mathbf{u}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u}, \mathbf{x})\mathrm{d}\mathbf{f}p(\mathbf{u})$$

$$\mathbf{f}, \mathbf{u} \mid \mathbf{x} \sim N\left(0, \begin{bmatrix} \mathbf{K}_{ff} & \mathbf{K}_{fu} \\ \mathbf{K}_{uf} & \mathbf{K}_{uu} \end{bmatrix}\right)$$

$$y_i | f_i \sim N(0, \sigma^2)$$

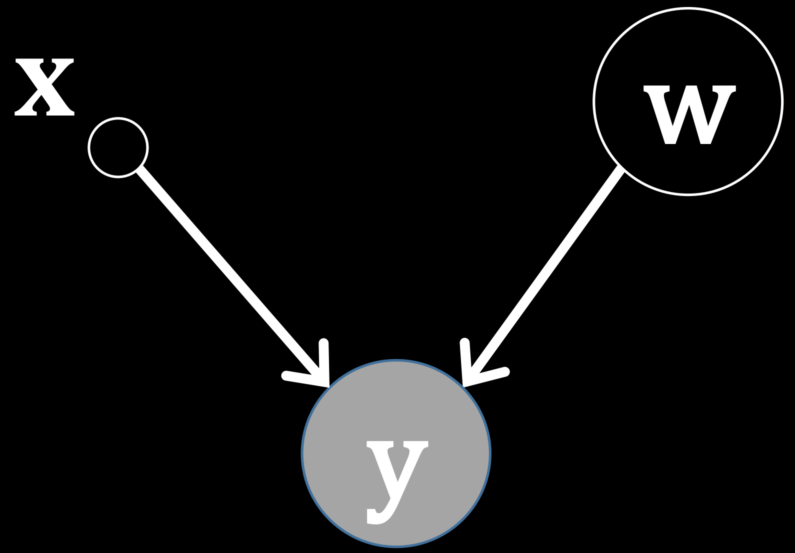
$$p(\mathbf{y}|\mathbf{u}) = N(\mathbf{y}|\mathbf{m}, \mathbf{C} + \sigma^2\mathbf{I})$$

$$\mathbf{C} = \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}$$

$$\mathbf{m} = \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}$$

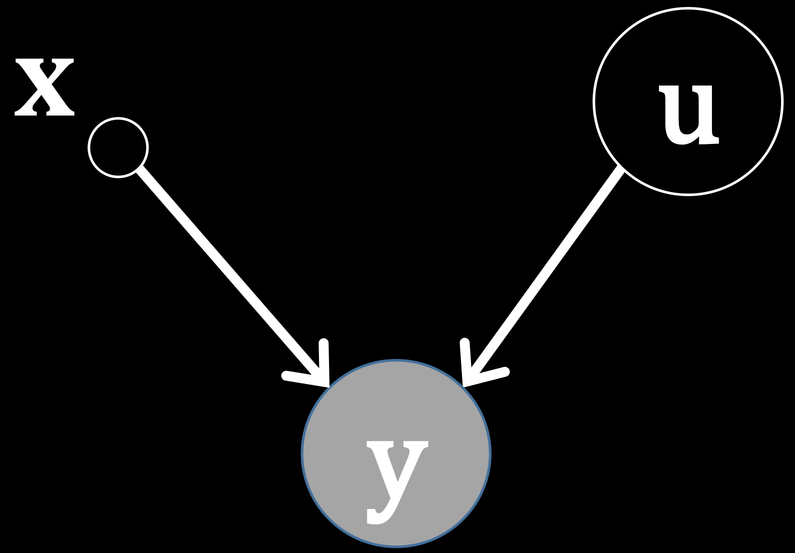
$$p(\mathbf{y}|\mathbf{u},\mathbf{x}) \geq \prod_{i1}^{n} \exp \int \ p(f_i|\mathbf{u},\mathbf{x}) \log p(y_i|f_i) \mathrm{d}\mathbf{f}$$

$$p(\mathbf{y}, \mathbf{w}|\mathbf{x}) = p(\mathbf{y}|\mathbf{w}, \mathbf{x})p(\mathbf{w})$$



$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{w}, \mathbf{x})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

$$p(\mathbf{y}, \mathbf{u}|\mathbf{x}) = p(\mathbf{y}|\mathbf{u}, \mathbf{x})p(\mathbf{u})$$



u looks like a parameter

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{u}, \mathbf{x})p(\mathbf{u})\mathrm{d}\mathbf{u}$$

but we can change the dimensionality of **u**

$$p(\mathbf{y}|\mathbf{u},\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf} + \sigma^2\mathbf{I})$$

$$p(\mathbf{y}|\mathbf{u},\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{m}, \mathbf{C} + \sigma^2\mathbf{I})$$

$$\mathbf{C} = \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}$$

$$\mathbf{m} = \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}$$

$$p(\mathbf{y}|\mathbf{u},\mathbf{x}) \geq \prod_{i1}^{n} \exp\langle \log p(y_i|f_i)\rangle_{p(f_i|\mathbf{u},\mathbf{x})}$$

$$\hat{p}(\mathbf{y}|\mathbf{u},\mathbf{x}) \geqq N(\mathbf{y}|\mathbf{m},\sigma^2\mathbf{I})\exp\left(\frac{c_{ii}}{2\sigma^2}\right)$$

$$c_{ii} = k_{ii} - \mathbf{k}_{iu}\mathbf{K}_{uu}^{-1}\mathbf{k}_{ui}$$

$$\mathbf{m} = \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}$$

model is linear in $\mathbf{u}$

expected
log likelihood

dissimilarity
between $q(\mathbf{x})$
and $p(\mathbf{x})$

$$\mathcal{L}(\mathbf{y}|\mathbf{u}) = \langle \log \hat{p}(\mathbf{y}|\mathbf{u}, \mathbf{x}) \rangle_{q(\mathbf{x})} - \mathrm{KL}(q(\mathbf{x})|p(\mathbf{x}))$$

model remains linear in $\mathbf{u}$

$$\hat{p}(\mathbf{y}|\mathbf{u},\mathbf{x}) \geqq N(\mathbf{y}|\mathbf{m},\sigma^2\mathbf{I})\exp\left(\frac{c_{ii}}{2\sigma^2}\right)$$

$$c_{ii} = k_{ii}(x_i, x_i) - \mathbf{k}_{iu}(x_i)\mathbf{K}_{uu}^{-1}\mathbf{k}_{ui}\,(x_i)$$

$$\mathbf{m}(\mathbf{x}) = \mathbf{K}_{fu}(\mathbf{x})\mathbf{K}_{uu}^{-1}\mathbf{u}$$

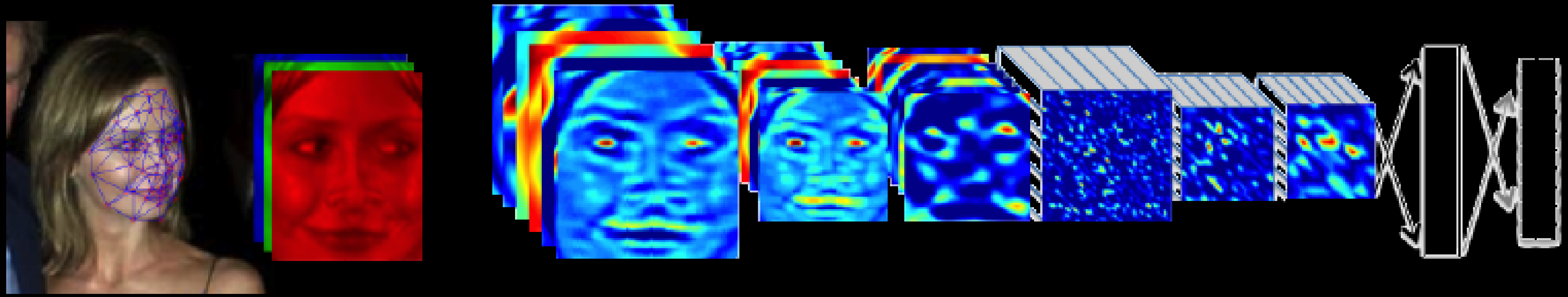
model is not linear in $\mathbf{x}$

$$\langle k_{ii}(x_{\boldsymbol{i}}, x_{\boldsymbol{i}}) \rangle_{q(x_i)}$$

$$\left\langle \mathbf{K}_{fu}(\mathbf{x}) \right\rangle_{q(\mathbf{x})}$$

$$\left\langle \mathbf{K}_{uf}(\mathbf{x}) \mathbf{K}_{fu}(\mathbf{x}) \right\rangle_{q(\mathbf{x})}$$

$$\mathbf{g}(x)$$



$\mathbf{f}_1(x)$ $\mathbf{f}_2(\cdot)$ $\mathbf{f}_3(\cdot)$ $\mathbf{f}_4(\cdot)$ $\mathbf{f}_5(\cdot)$ $\mathbf{f}_6(\cdot)$ $\mathbf{f}_7(\cdot)$ $\mathbf{f}_8(\cdot)$ $\mathbf{f}_9(\cdot)$

$$\mathbf{g}(x) = \mathbf{f}_9\left(\mathbf{f}_8\left(\mathbf{f}_7(\mathbf{f}_6(\cdots))\right)\right)$$

two Gaussian processes: apply bound recursively

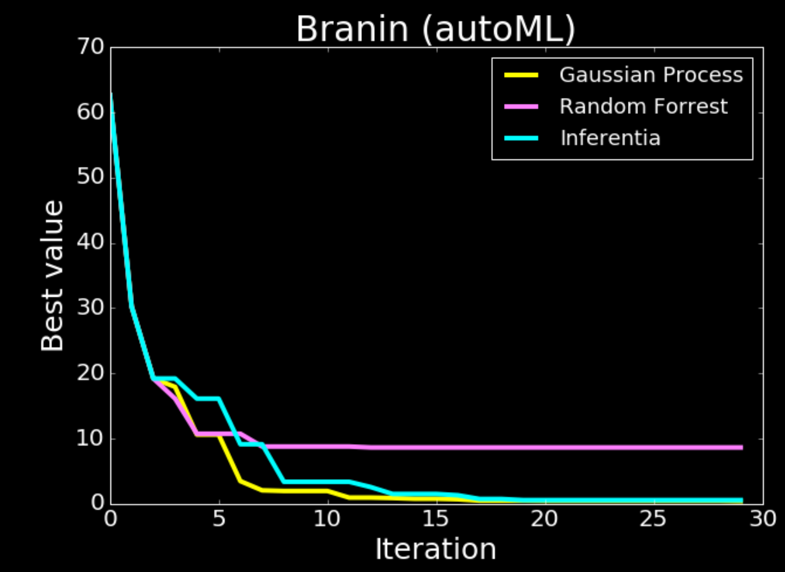
$$\int p(y|\mathbf{f}_5)p(\mathbf{f}_5|\mathbf{f}_4)p(\mathbf{f}_4|\mathbf{f}_3)p(\mathbf{f}_3|\mathbf{f}_2)p(\mathbf{f}_1|\mathbf{x})\mathrm{d}\mathbf{f}$$

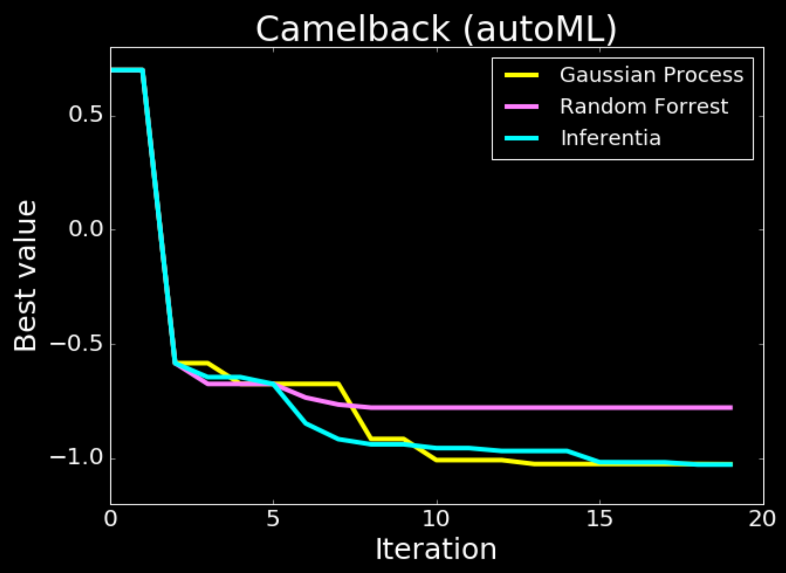
$$\mathbf{g}(x) = \mathbf{f}_5\left(\mathbf{f}_4\left(\mathbf{f}_3\left(\mathbf{f}_2(\mathbf{f}_1(x))\right)\right)\right)$$

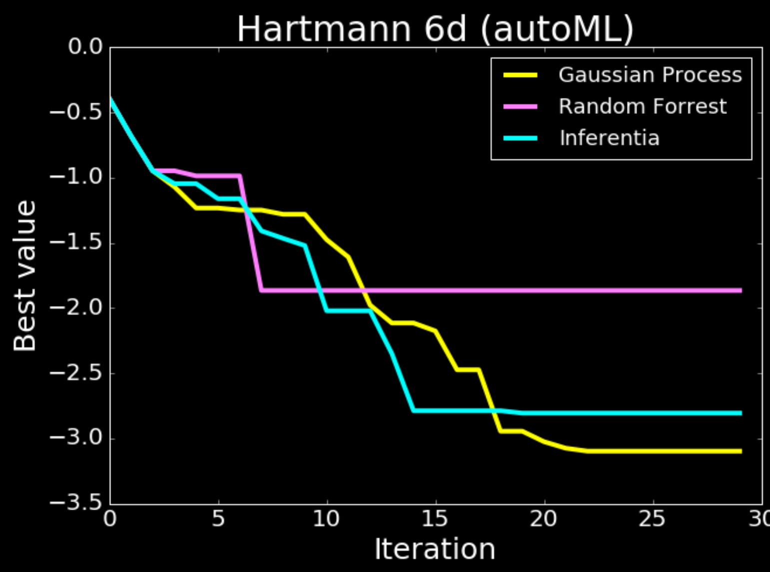
Regression

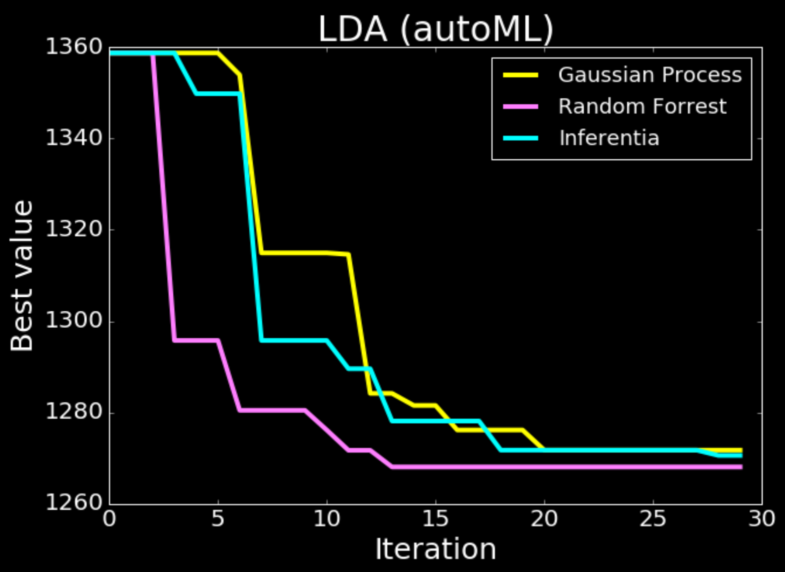
data set	n	p	GP	Sparse GP	Deep GP
housing	506	13	2.78±0.54	2.77±0.60	2.69±0.49
redwine	588	11	0.72±0.06	0.62±0.04	0.62±0.04
energy1	768	8	0.48±0.07	0.50±0.07	0.49±0.07
energy2	768	8	0.59±0.08	1.66±0.21	1.39±0.49
concrete	1030	8	5.26±0.67	5.81±0.62	5.66±0.62

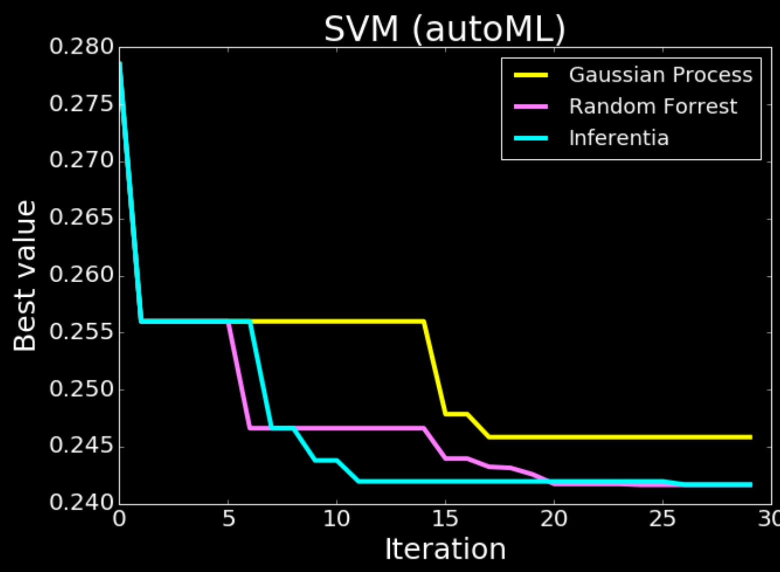
Bayesian Optimization



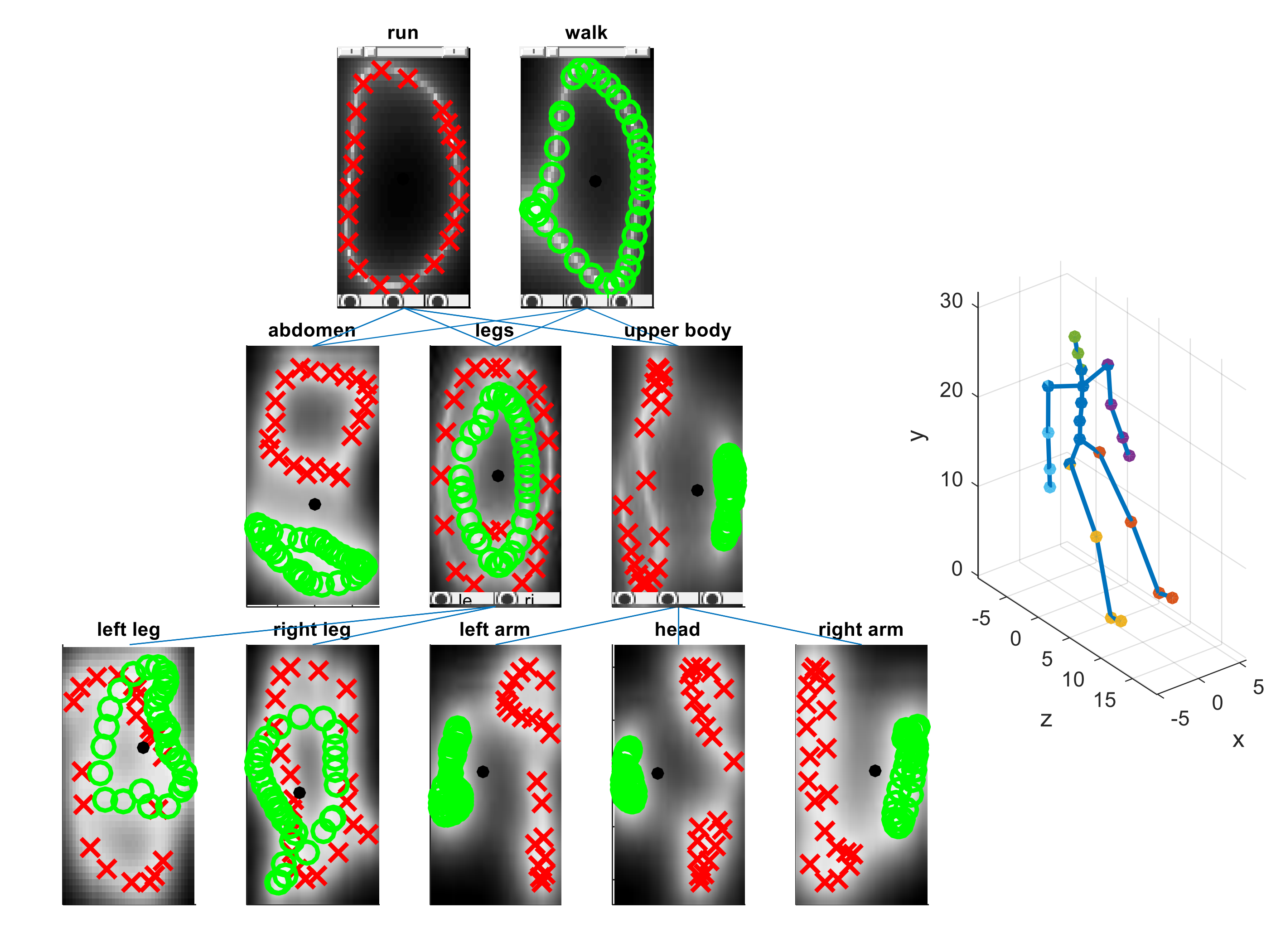




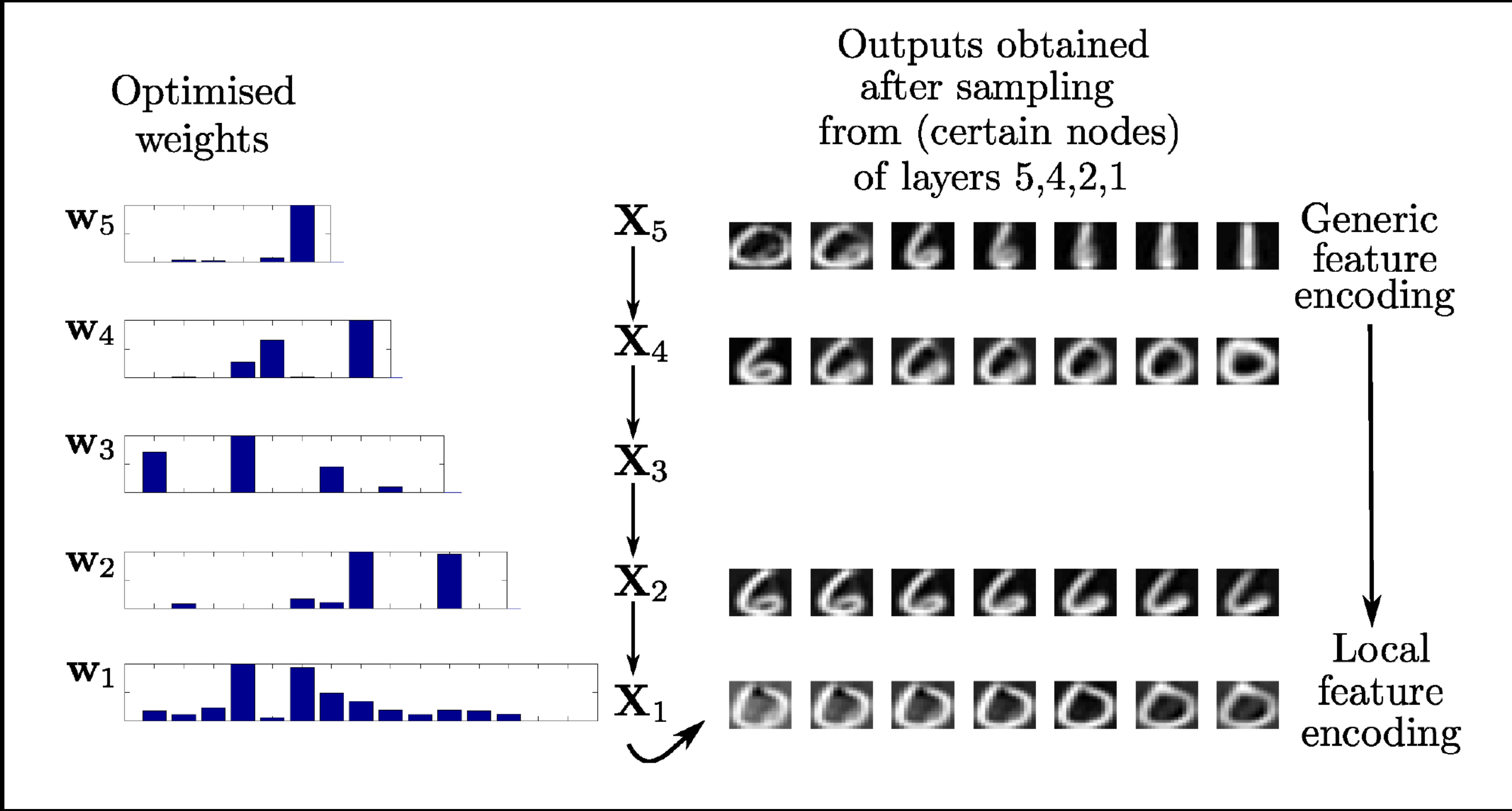




Example: Motion Capture Modelling



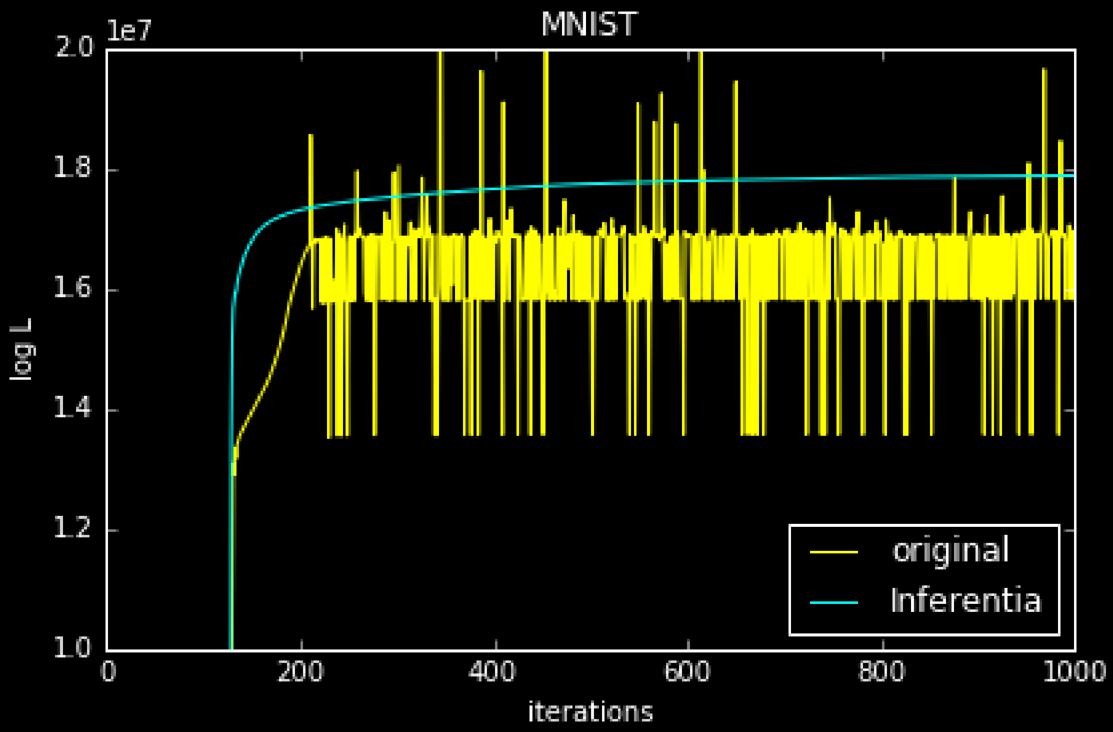
Modelling Digits

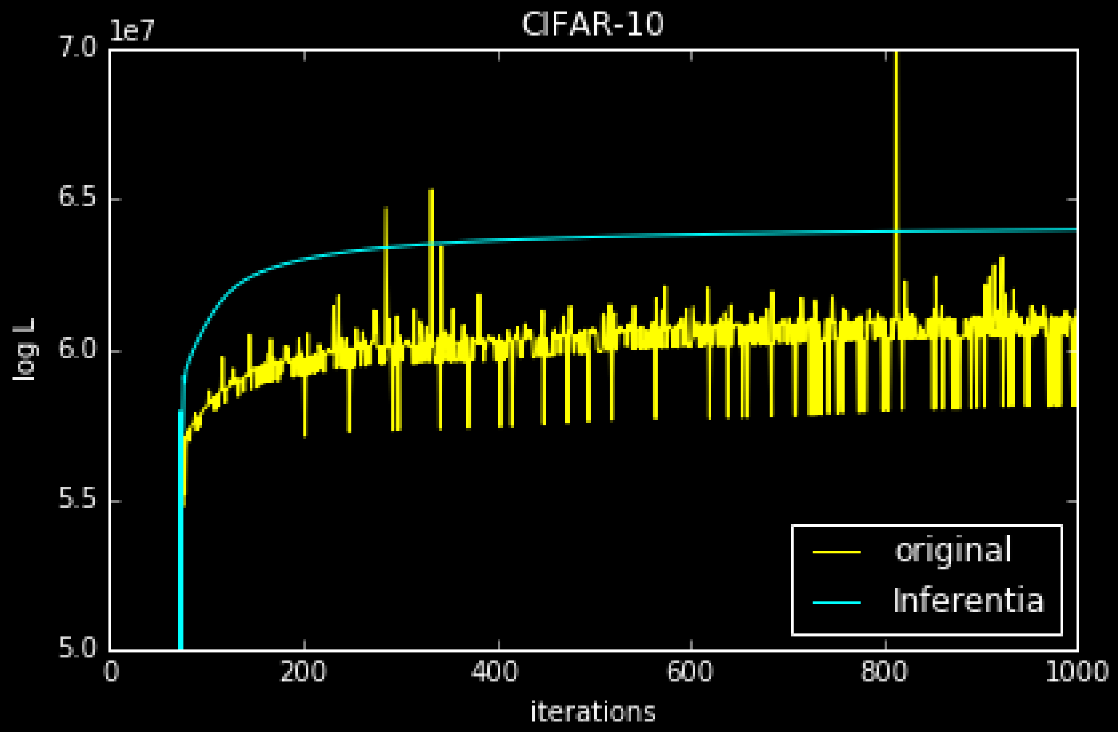


Inferentia

Challenging Uncertainty

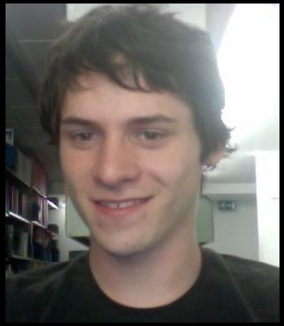
Numerical Issues



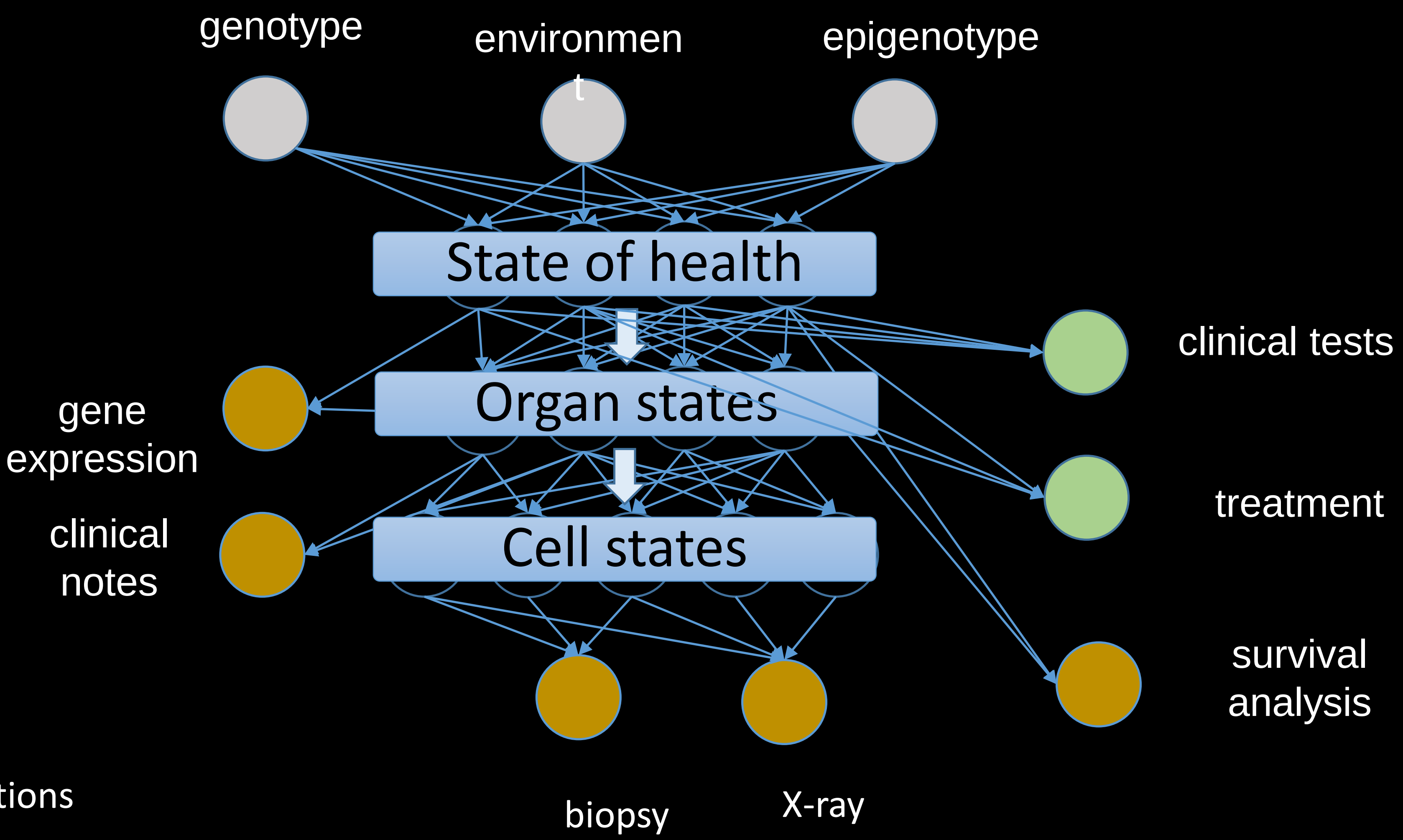


Health





- Complex system
- Scarce data
- Different modalities
- Poor understanding of mechanism
- Large scale



PLoS Comp Bio, Nature Communications

Thank you

Neil Lawrence
http://inverseprobability.com
@lawrennd