Beyond Backpropagation: Uncertainty Propagation

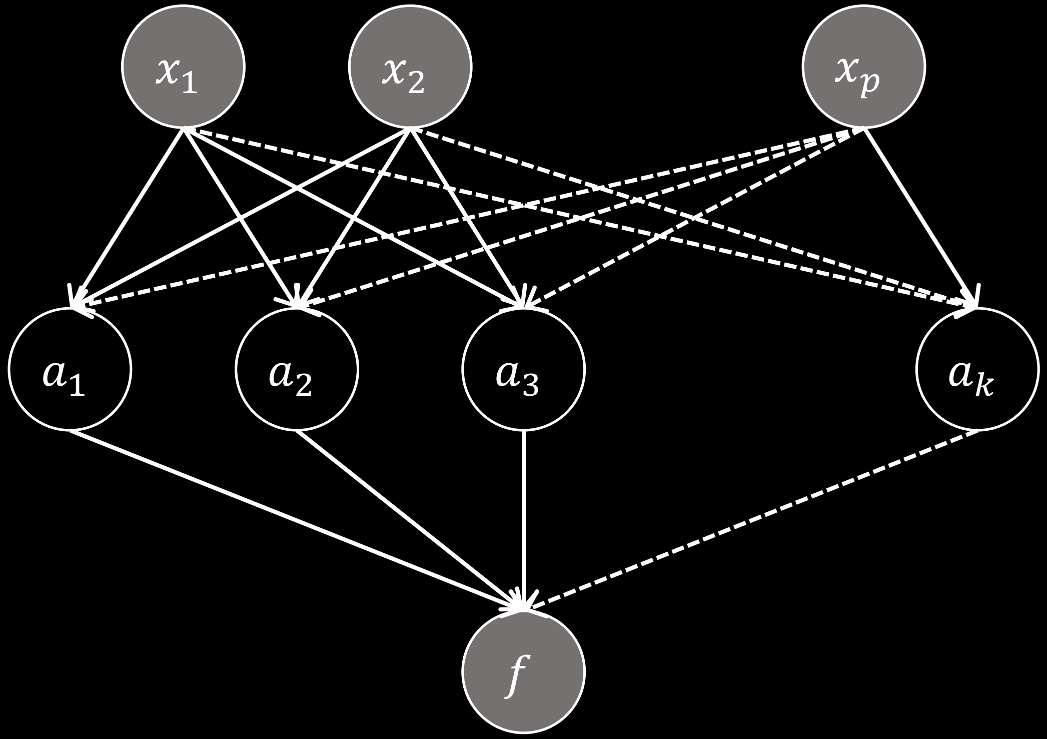
NEIL LAWRENCE

UNIVERSITY OF SHEFFIELD

@lawrennd

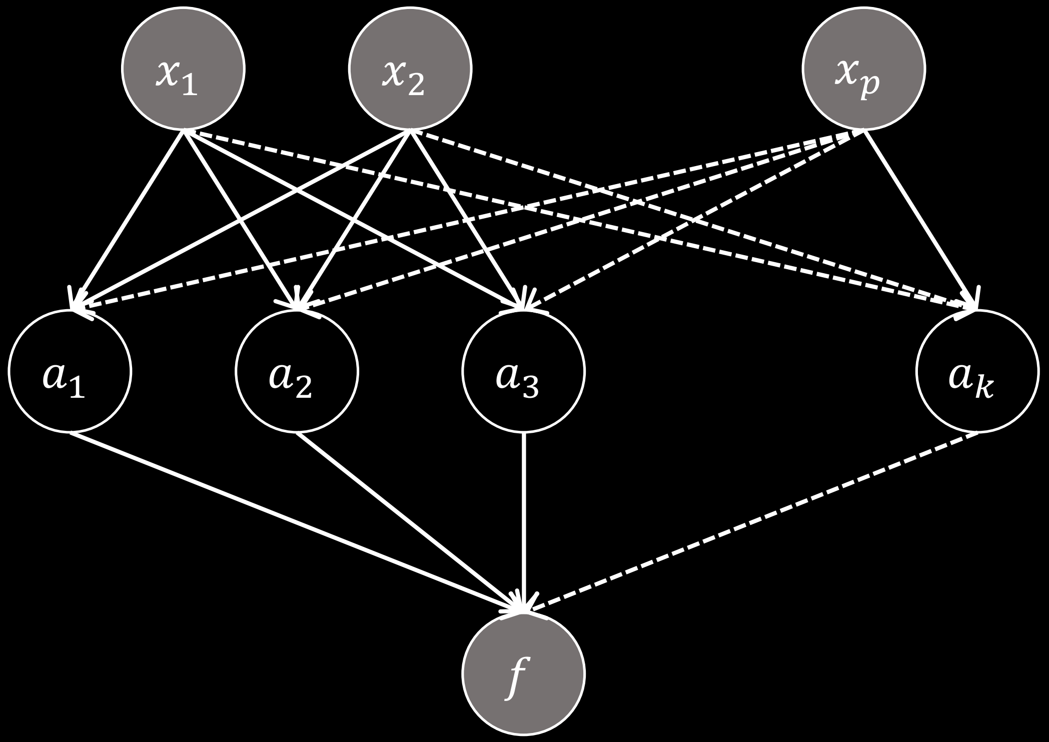
$$f(\mathbf{x}) = \sum_{j=1}^{k} u_j \phi(a_j)$$

$$a_j = \sum_{i=1}^{p} v_{i,j}\, x_i$$



$$v_{i,j} \sim N(0, \alpha_u)$$

$$u_i \sim N(0, \alpha_u)$$



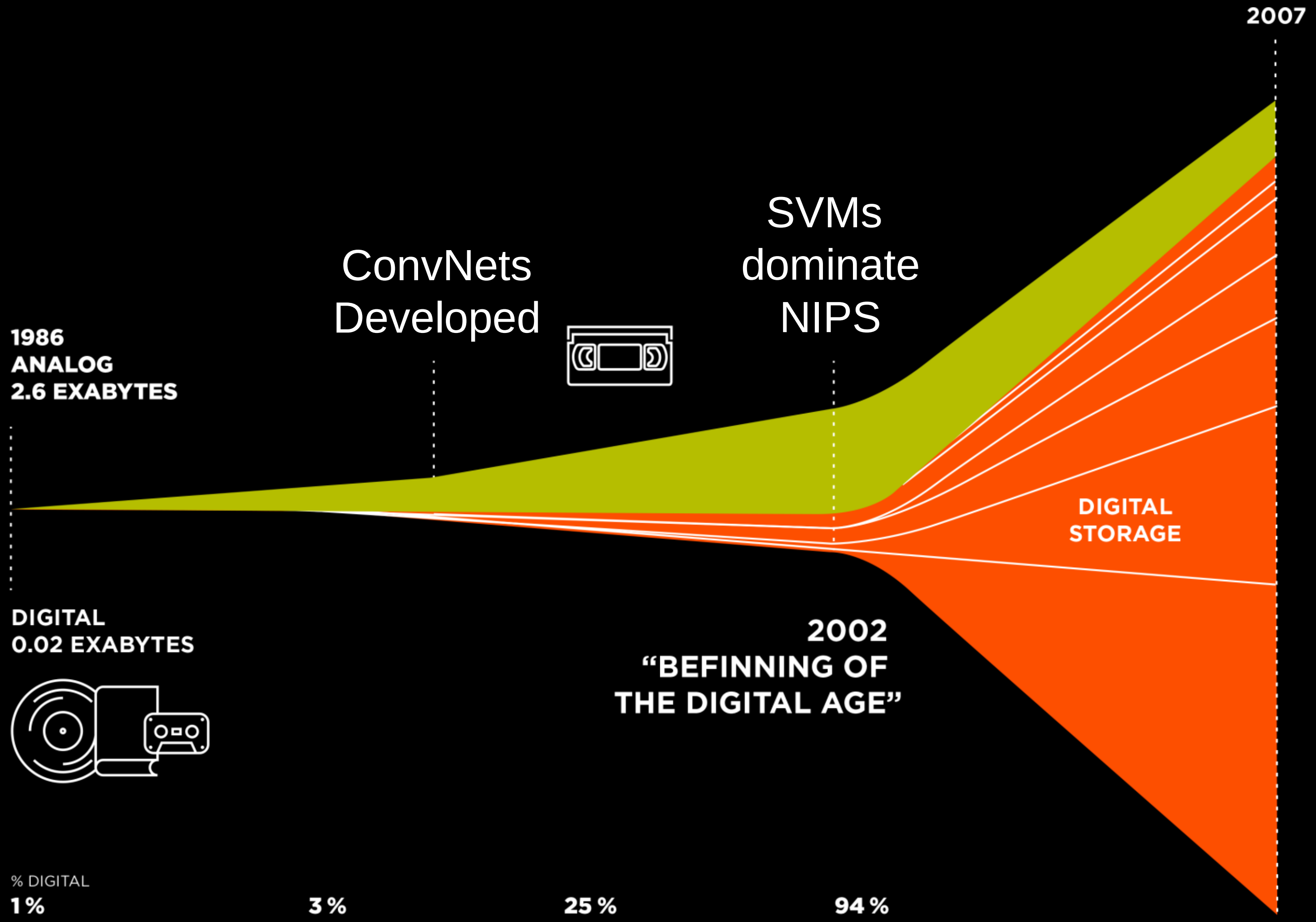
$$E(\mathbf{u},\mathbf{V}) = \sum_{i=1}^{n} \left(y_i - f_i(\mathbf{x}_i;\mathbf{u},\mathbf{V})\right)^2$$

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

$$\log p(\mathbf{y}|\mathbf{x},\mathbf{u},\mathbf{V}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y_i - f(\mathbf{x}_i;\mathbf{u},\mathbf{V})\big)^2 - \frac{1}{2\alpha_u}u_i^2 - \frac{1}{2\alpha_v}v_{i,j}^2$$

$$+ \text{const.}$$

GLOBAL INFORMATION STORAGE CAPACITY
IN OPTIMALLY COMPRESSED BYTES



ANALOG
19 EXABYTES

- Paper, film, audiotape and vinyl: 6%
- Analog videotapes (VHS, etc): 94%

ANALOG ▲

DIGITAL ▼

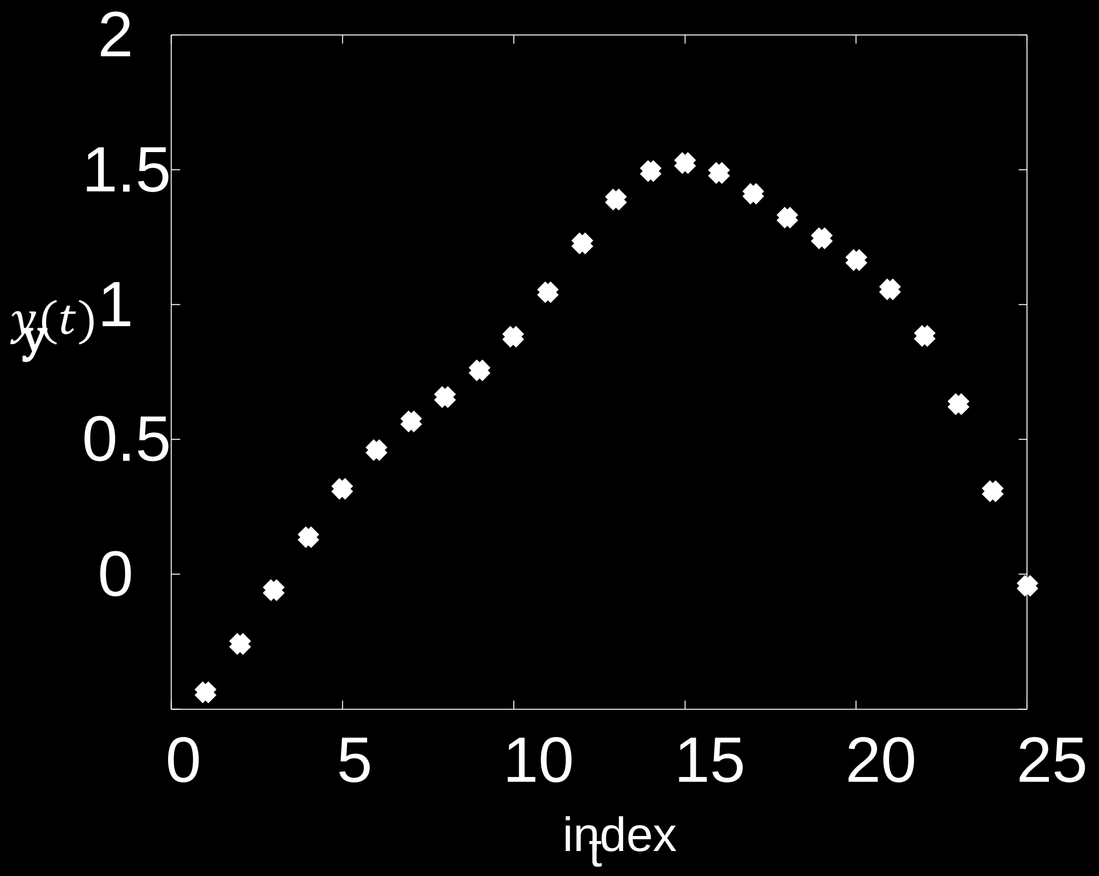
- Portable media, flash drives: 2%
- Portable hard disks: 2.4%
- CDs & Minidisks: 6.8%
- Computer Servers and Mainframes: 8.9%
- Digital Tape: 11.8%
- DVD/Blu-Ray: 22.8%
- PC Hard Disks: 44.5%
  123 Billion Gigabytes
- Others: < 1% (incl. Chip Cards, Memory Cards, Floppy Disks, Mobile Phones, PDAs, Cameras/Camcorders, Video Games)

DIGITAL
280 EXABYTES

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

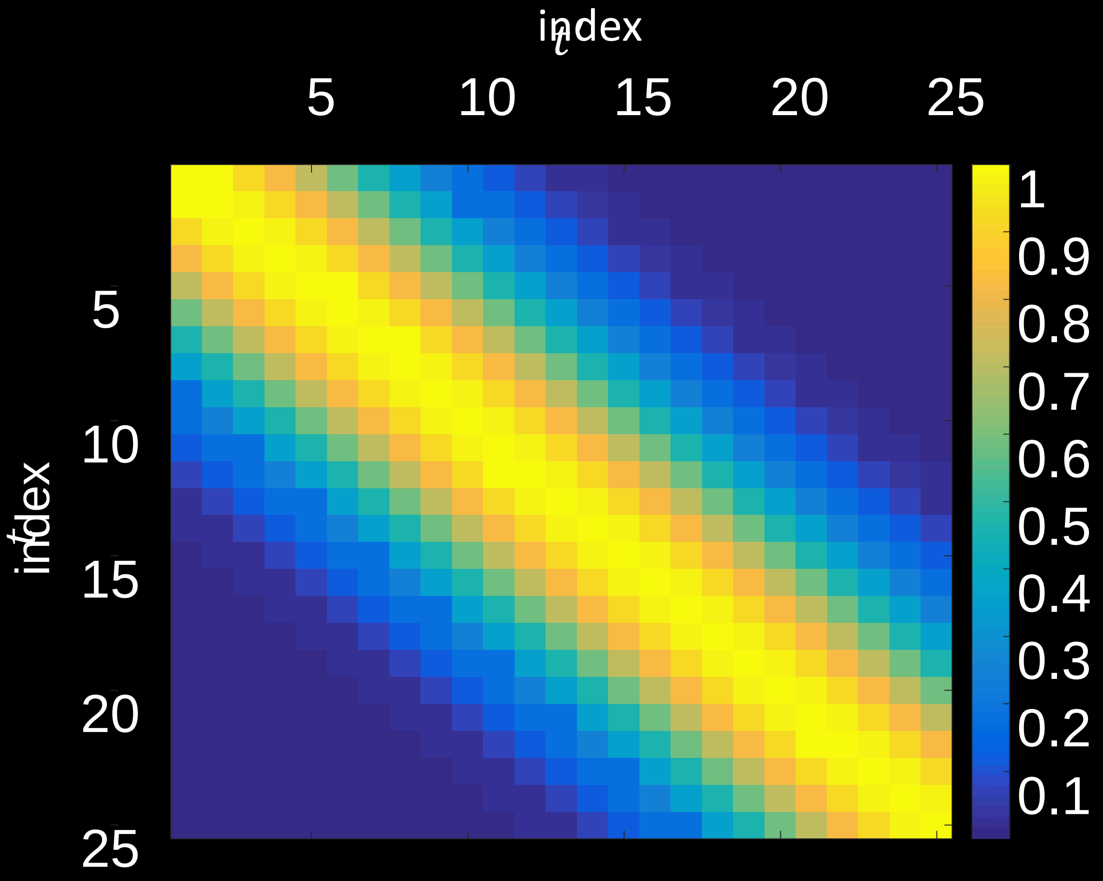
Zero Mean Gaussian Samples

Zero Mean Gaussian Process Sample



samples from Gaussian

samples from Gaussian process



covariance $\mathbf{C}$

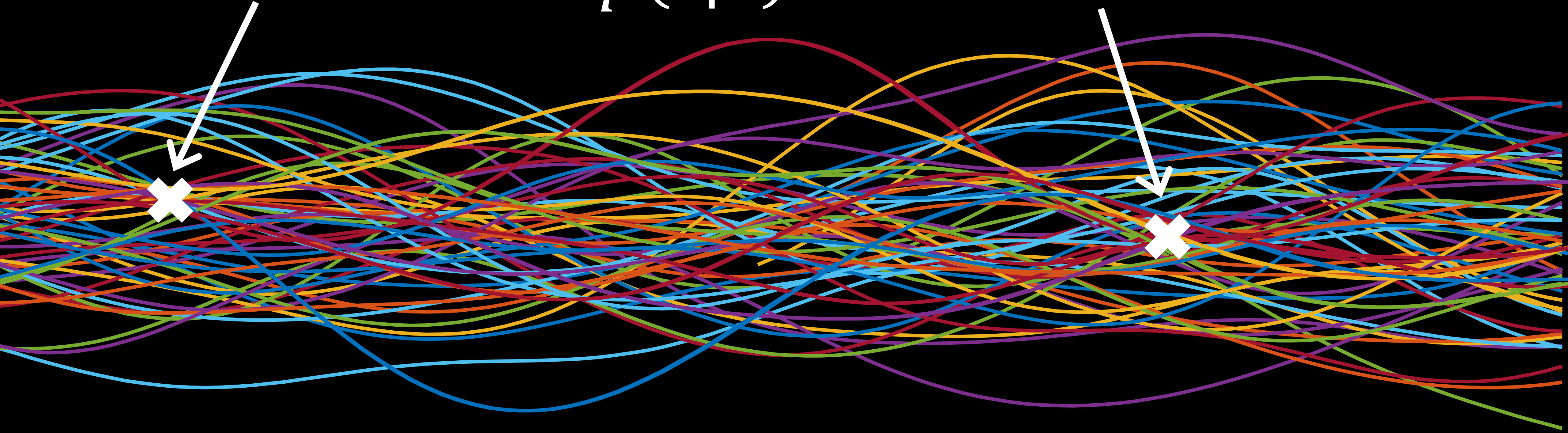
covariance function $c(t, t')$

Gaussian Processes

x_1, y_1

$p(\mathbf{f}|\mathbf{x})$

x_2, y_2



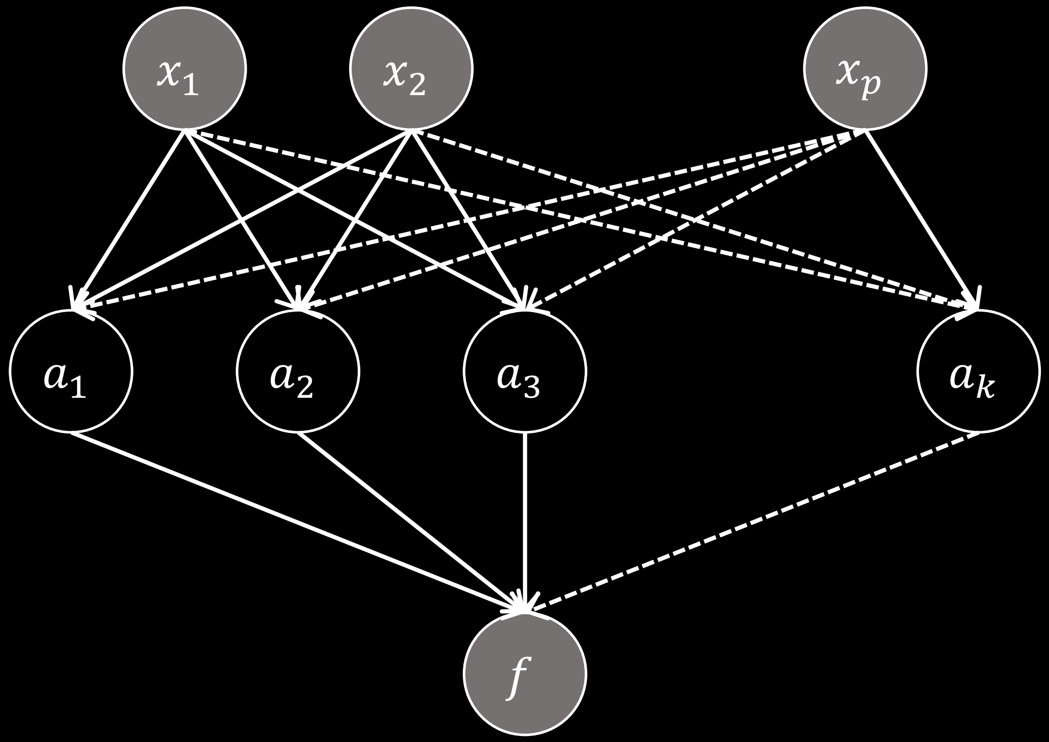
$p(y_1|f_1)$

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

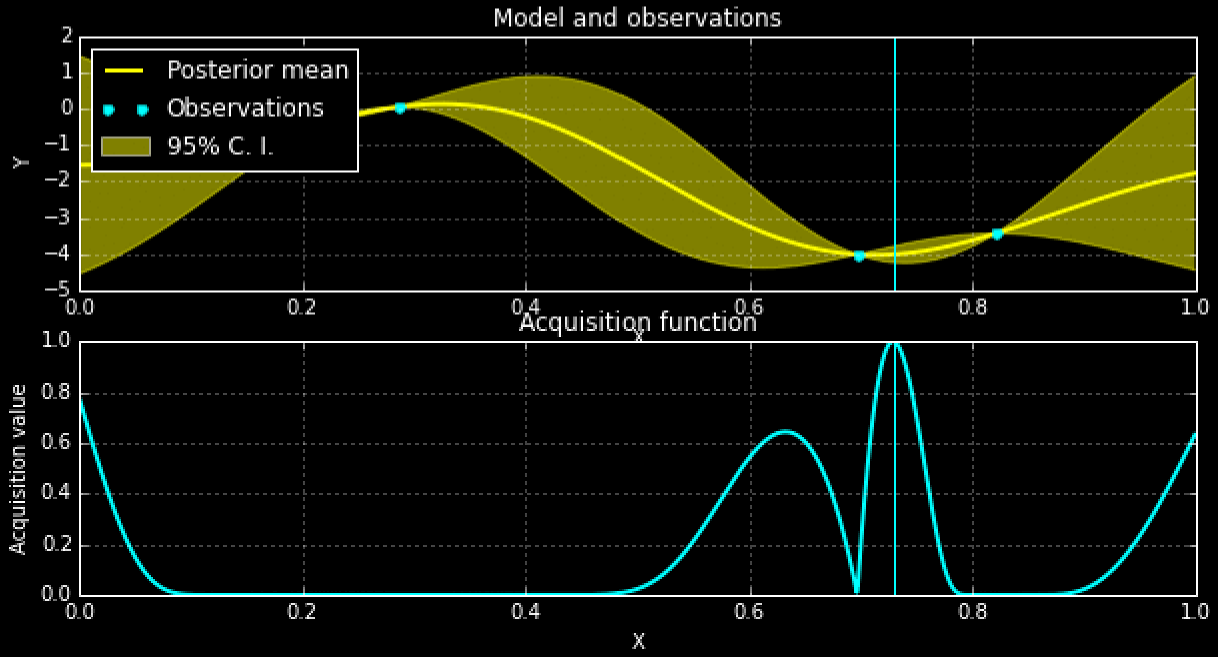
$p(y_2|f_2)$

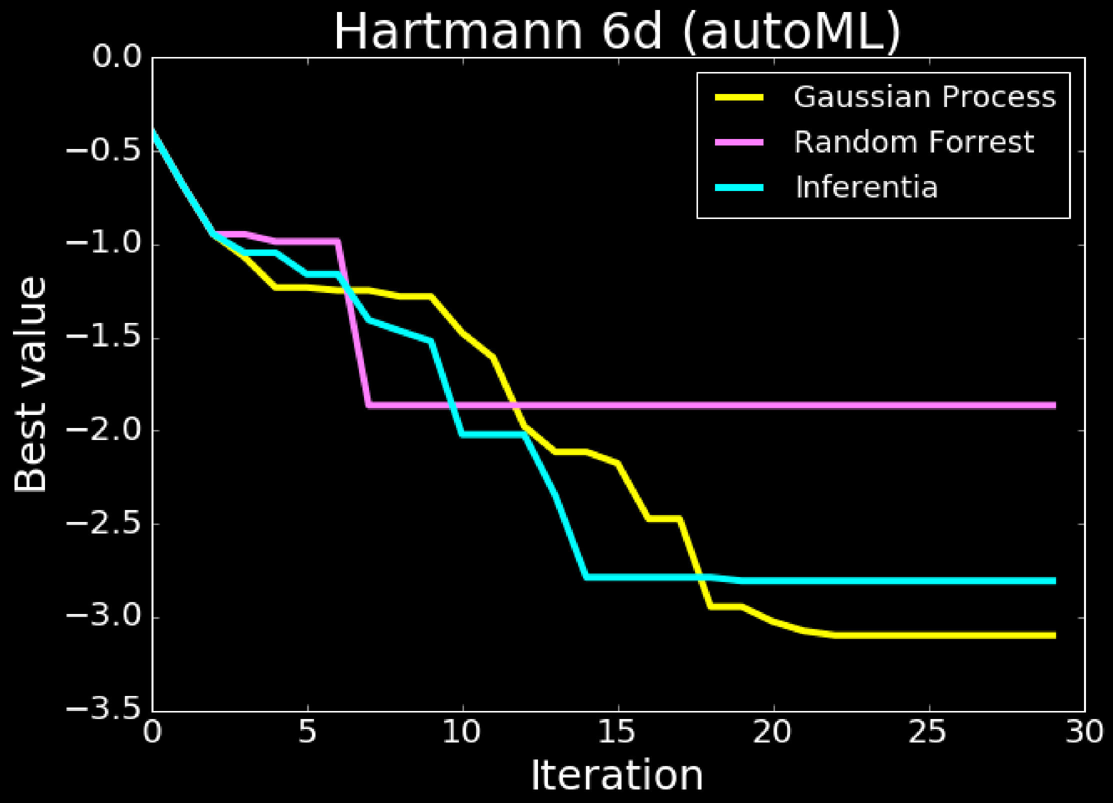
$$v_{i,j} \sim ?$$

$$u_i \sim N(0, \alpha_u)$$

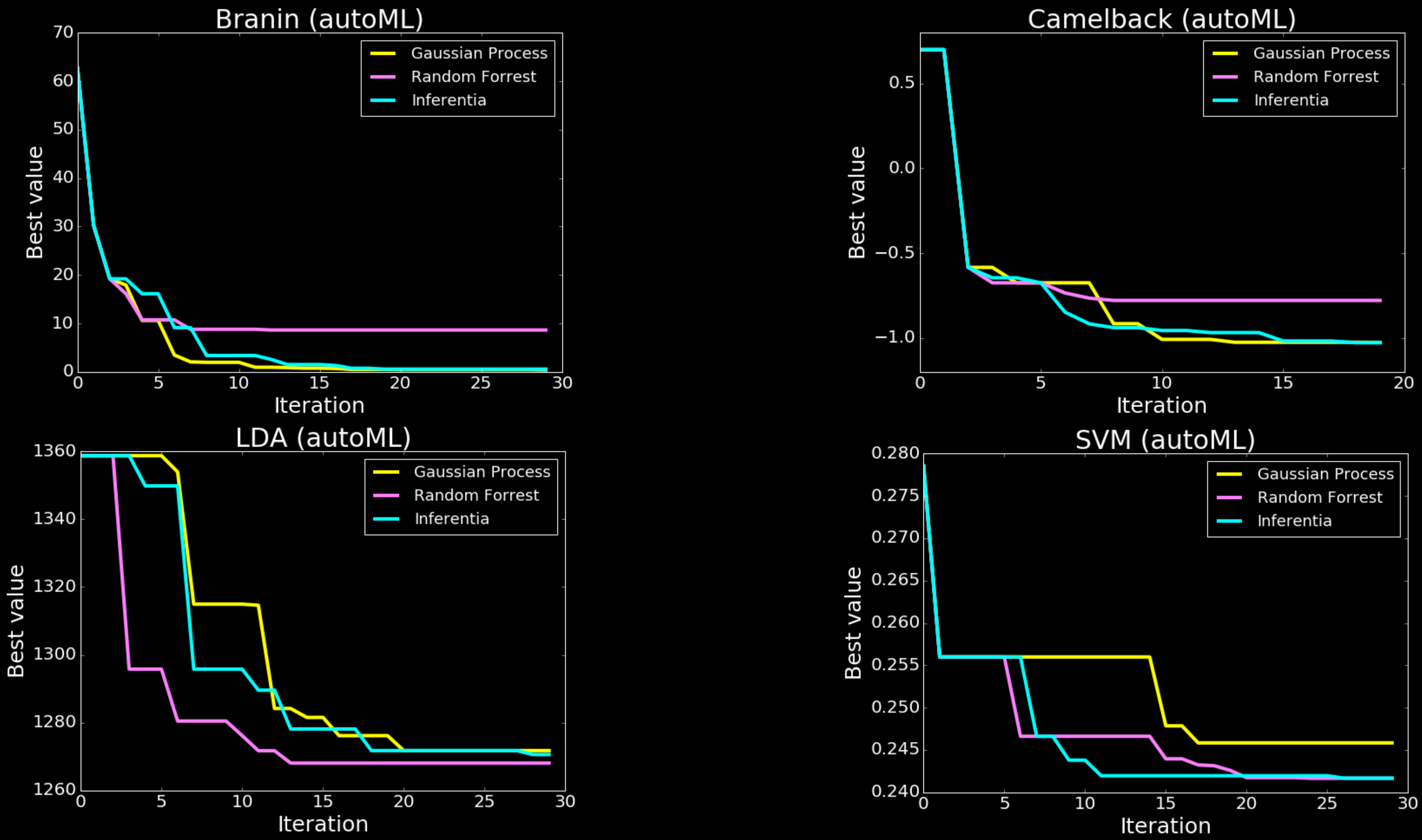


Bayesian Optimization





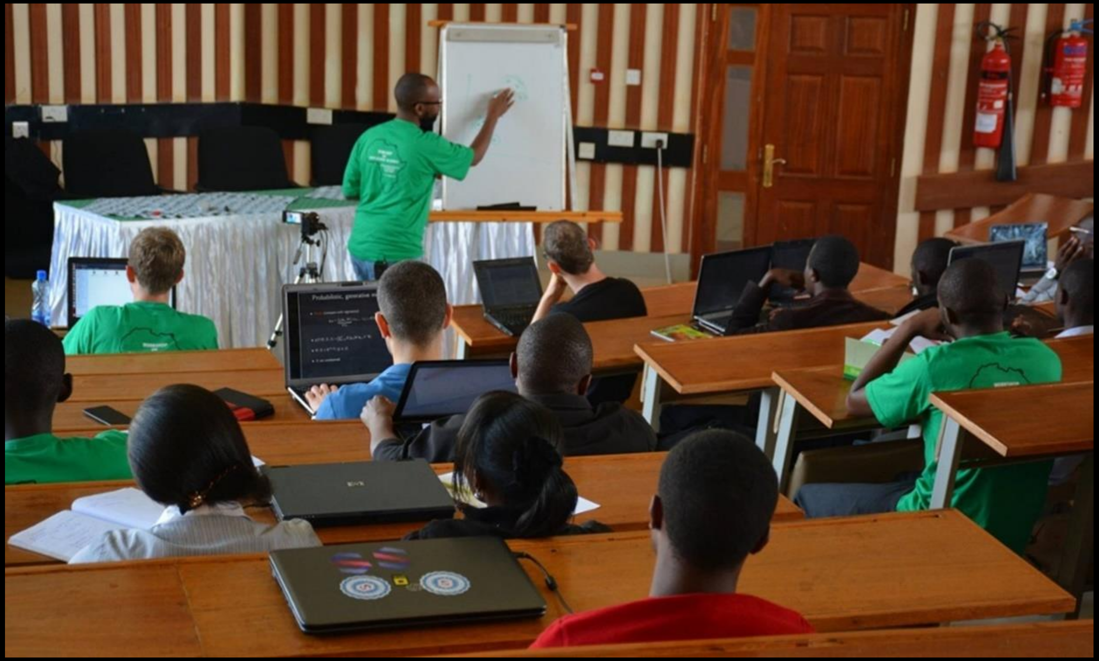
- Check http://sheffieldml.github.io/GPyOpt/


Branin (autoML)
Best value
Iteration
0
10
20
30
40
50
60
70
0
5
10
15
20
25
30
Gaussian Process
Random Forrest
Inferentia
Camelback (autoML)
Best value
Iteration
0.5
0.0
−0.5
−1.0
0
5
10
15
20
Gaussian Process
Random Forrest
Inferentia
LDA (autoML)
Best value
Iteration
1260
1280
1300
1320
1340
1360
0
5
10
15
20
25
30
Gaussian Process
Random Forrest
Inferentia
SVM (autoML)
Best value
Iteration
0.240
0.245
0.250
0.255
0.260
0.265
0.270
0.275
0.280
0
5
10
15
20
25
30
Gaussian Process
Random Forrest
Inferentia

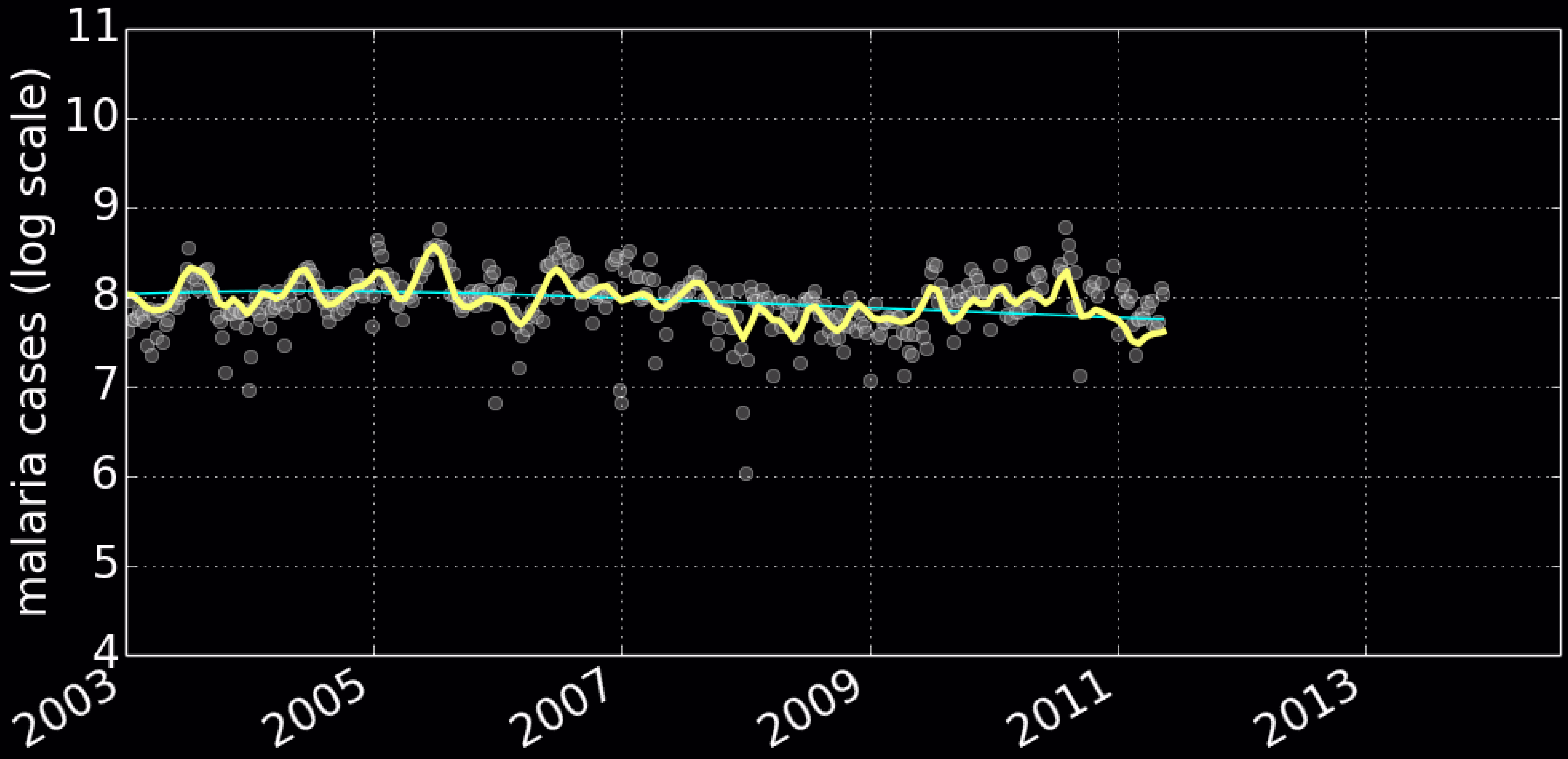
Open Data Science and Africa

Challenge

- “Whole pipeline challenge”
- Make software available
- Teach summer schools
- Support local meetings
    - Publicity in the Guardian
- Opportunities to deploy pipeline solutions

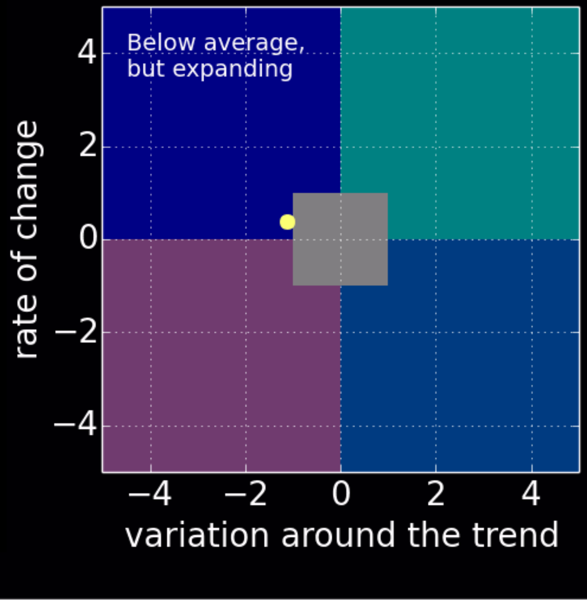


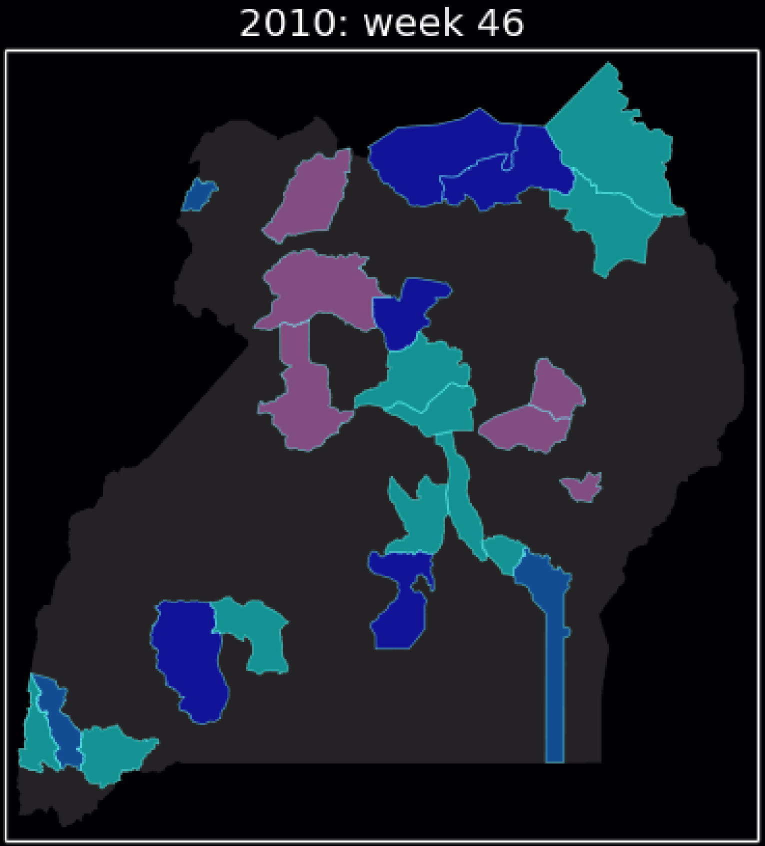
Disease Incidence for Malaria

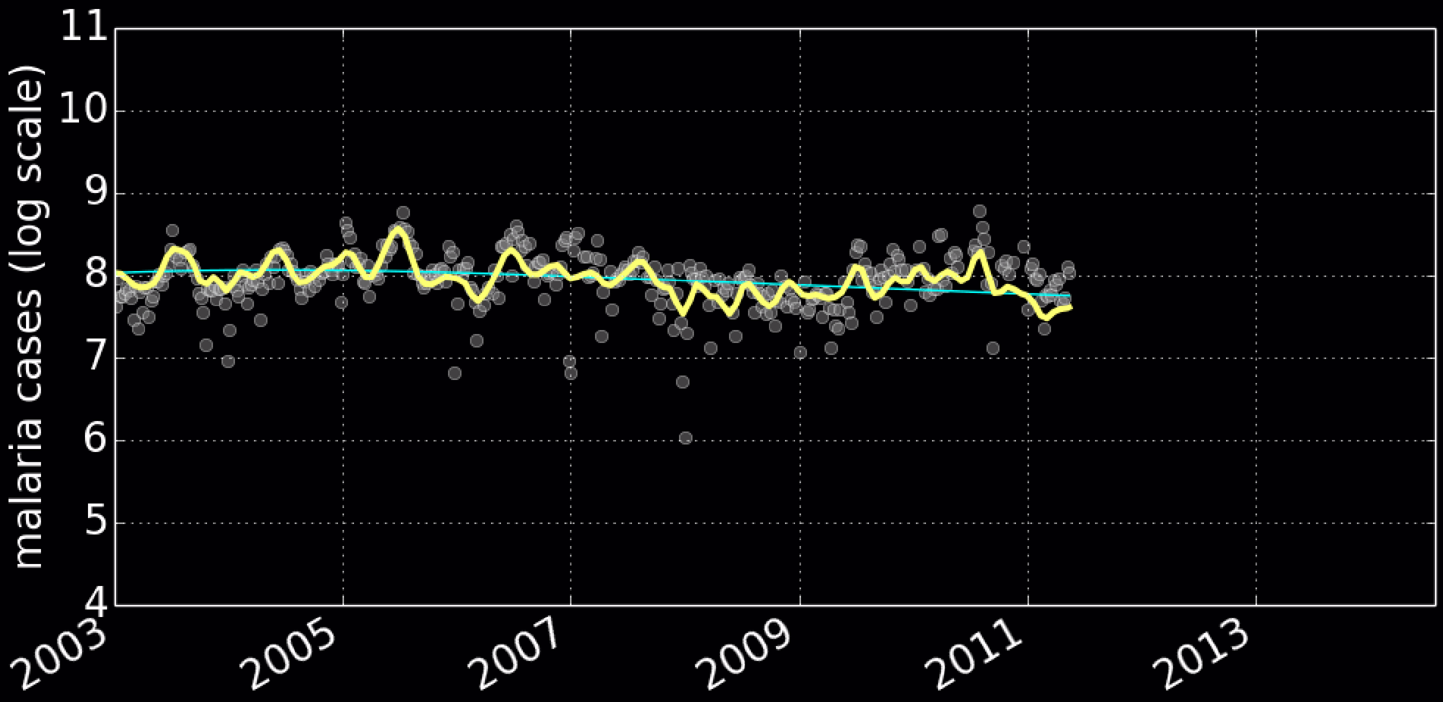


Uganda

- Spatial models of disease







Deployed with UN Global Pulse Lab

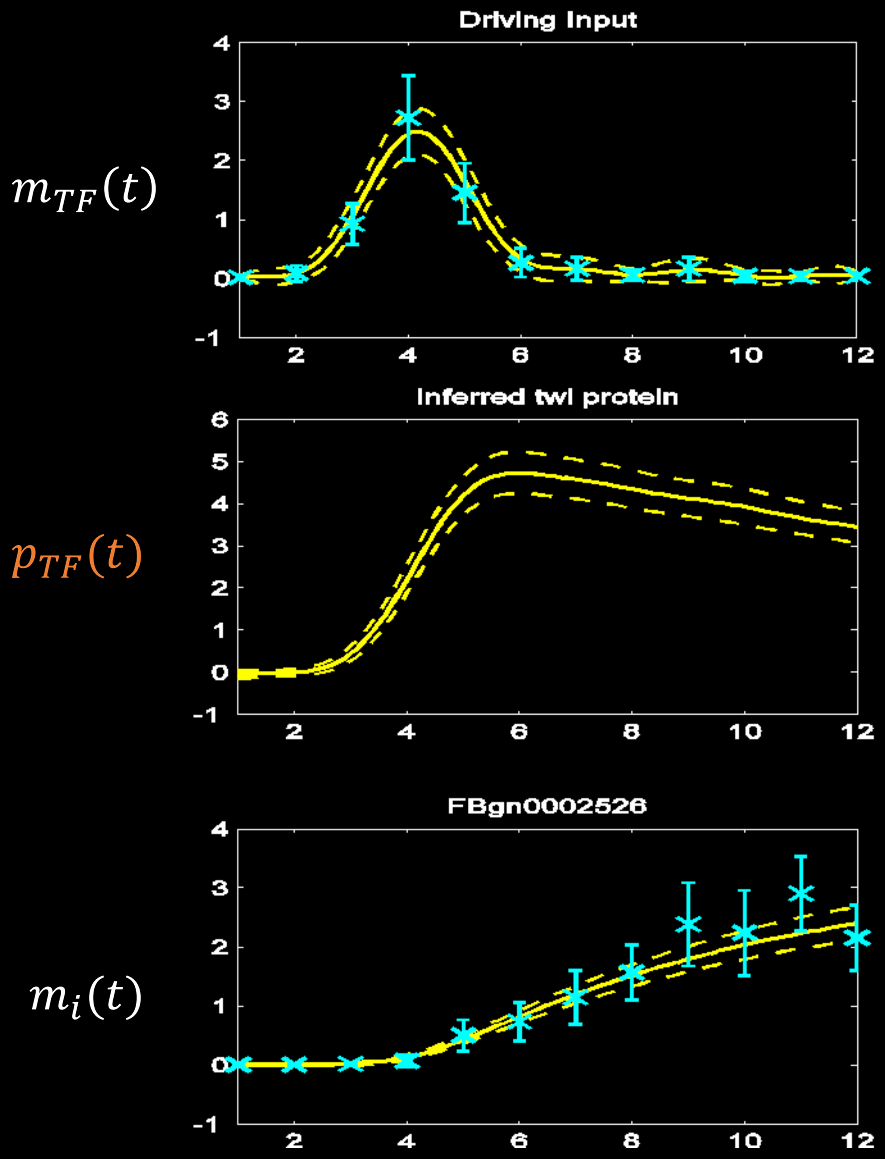
GLOBAL PULSE

http://pulselabkampala.ug/hmis/

Results

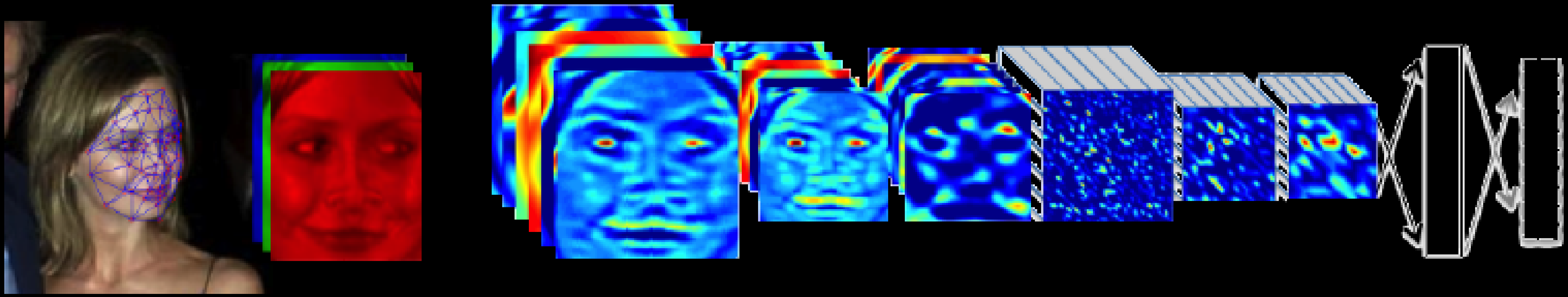
$$\frac{\mathrm{d}p_{TF}(t)}{\mathrm{d}t} = s_f m_{TF}(t) - d_f p_{TF}(t)$$

$$\frac{\mathrm{d}m_i(t)}{\mathrm{d}t} = s_i p_{TF}(t) - d_i m_i(t)$$



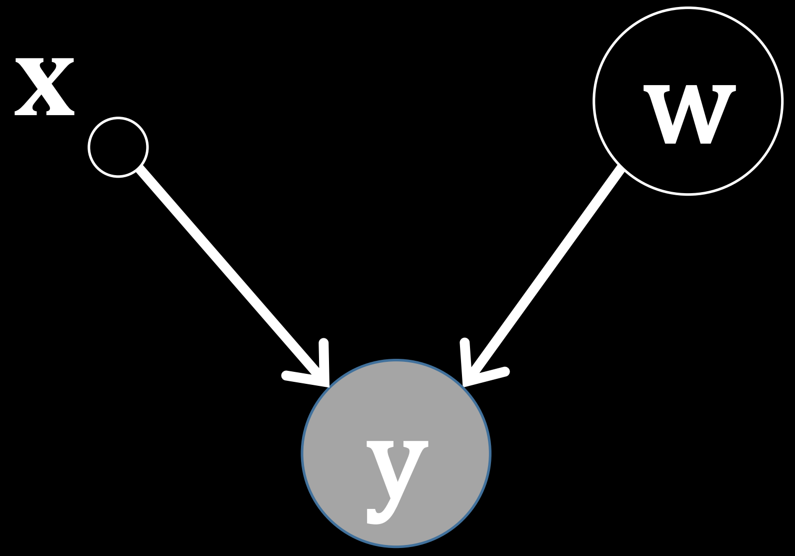
David: Have we thrown out the baby with the bathwater?

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



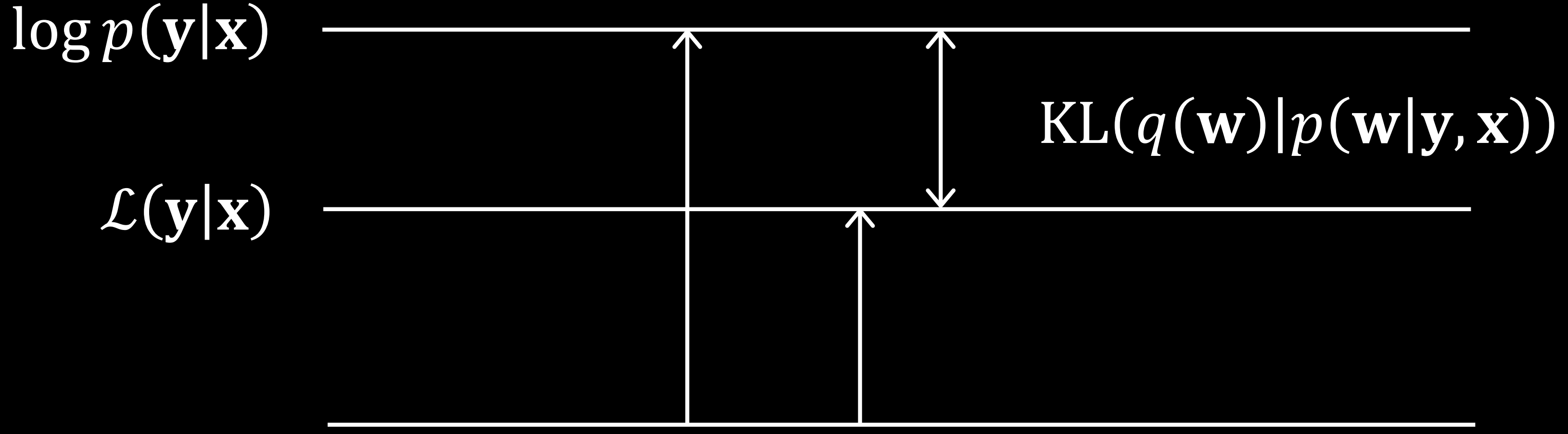
$$\mathbf{g}(x) = \mathbf{f}_9\left(\mathbf{f}_8\left(\mathbf{f}_7(\mathbf{f}_6(\cdots))\right)\right)$$

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



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

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



$\log p(\mathbf{y}|\mathbf{x})$

$\mathrm{KL}(q(\mathbf{w})|p(\mathbf{w}|\mathbf{y},\mathbf{x}))$

$\mathcal{L}(\mathbf{y}|\mathbf{x})$

expected
log likelihood

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

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

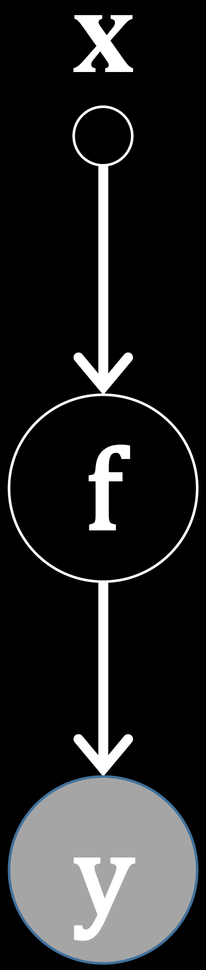
$$\mathcal{L}(\mathbf{y}|\mathbf{x}) = \left\langle \sum_{i=1}^{n} \left(y_i - g(\mathbf{x}_i;\mathbf{w})\right)^2 \right\rangle_{q(\mathbf{w})} - \mathrm{KL}(q(\mathbf{w})|p(\mathbf{w})) + \mathrm{const}$$

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

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

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

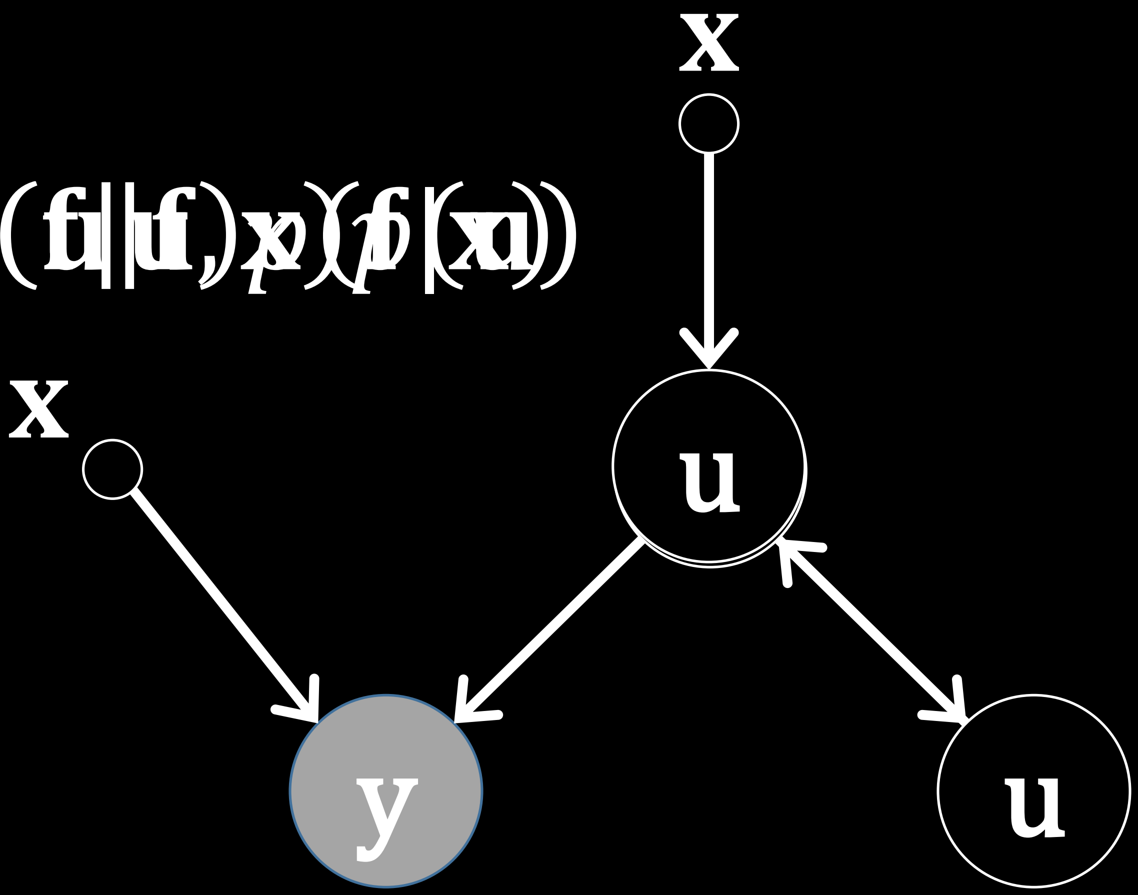
$$p(\mathbf{y}, \mathbf{f}|\mathbf{x}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x})$$



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

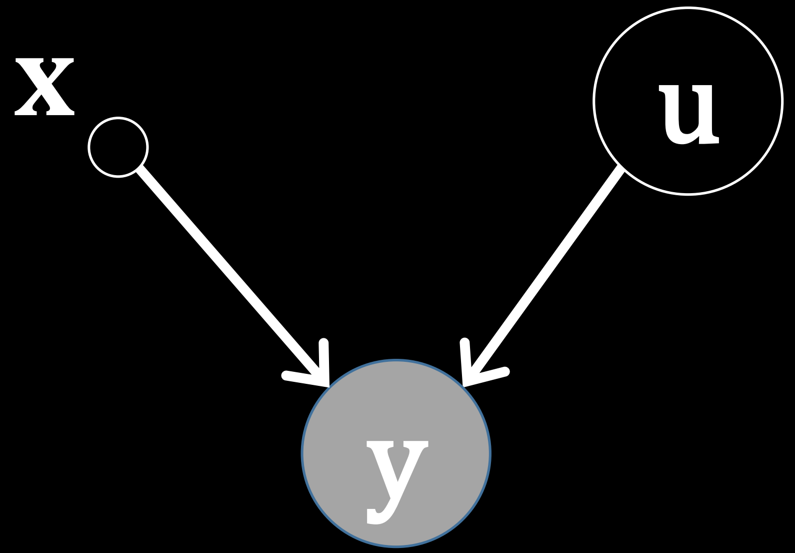
$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u}, \mathbf{x})p(\mathbf{u})$$

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



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

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



$\mathbf{u}$ looks like a parameter

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

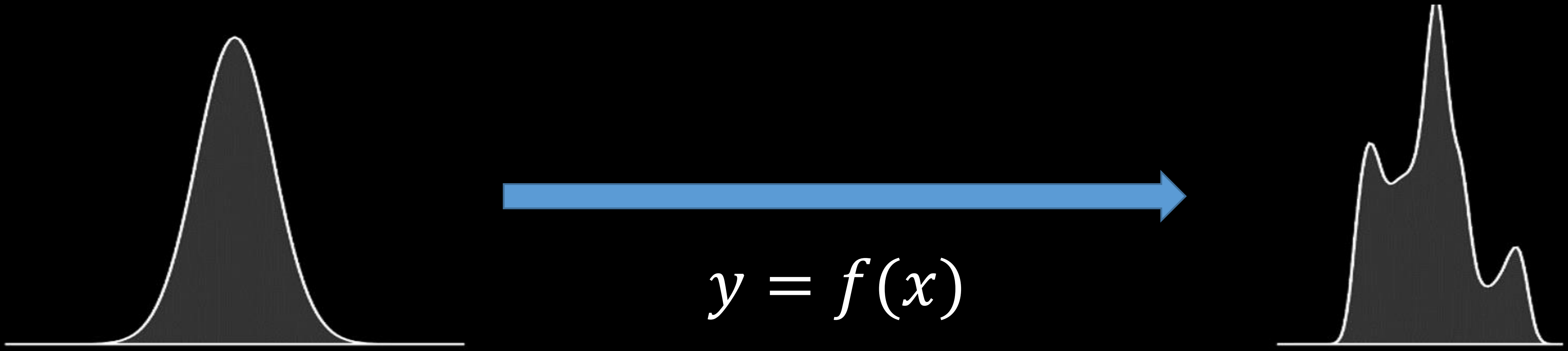
but we can change the dimensionality of $\mathbf{u}$

two Gaussian processes: apply bound recursively

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

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

Render Gaussian Non Gaussian



$$y = f(x)$$

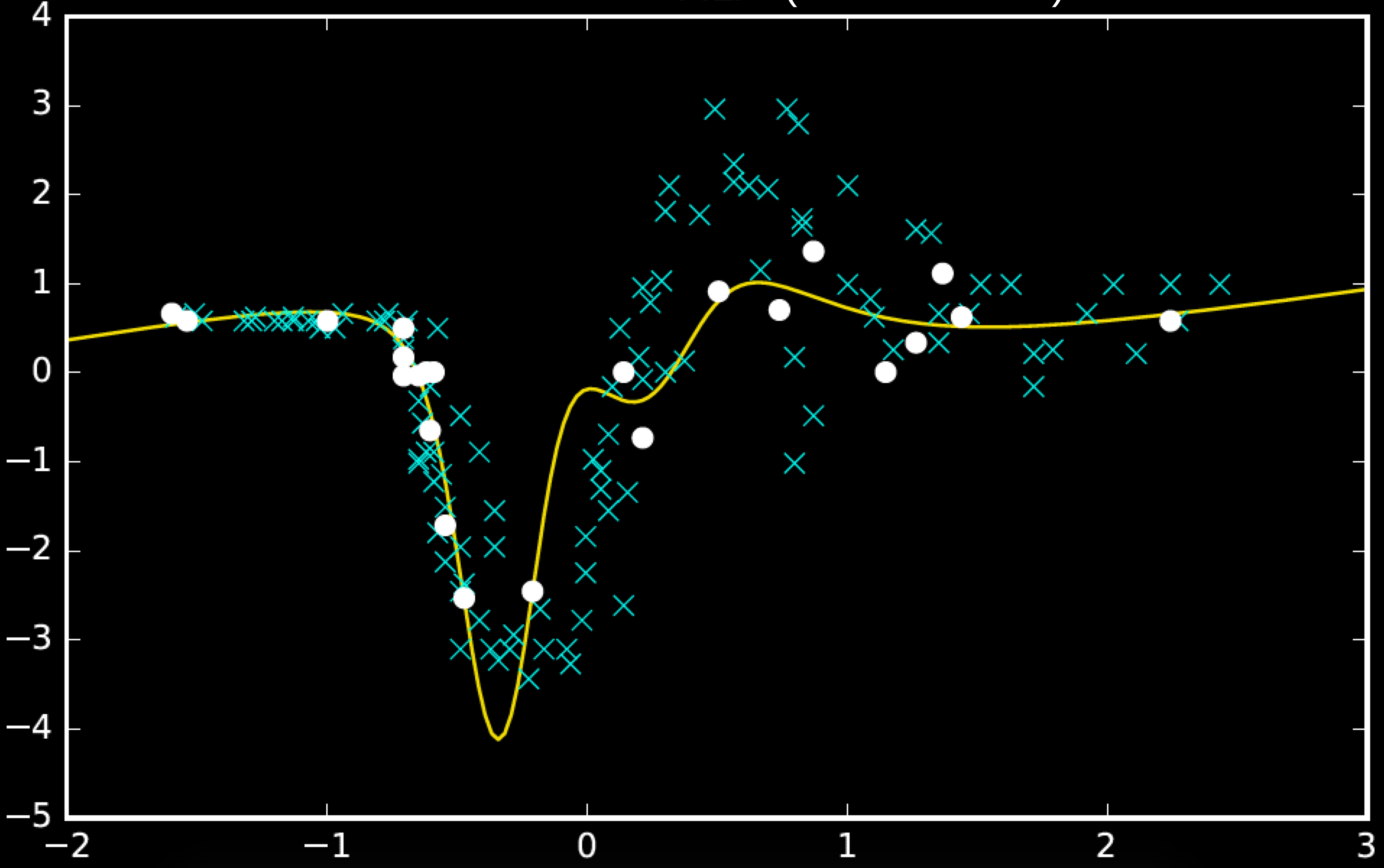
Stochastic Process Composition

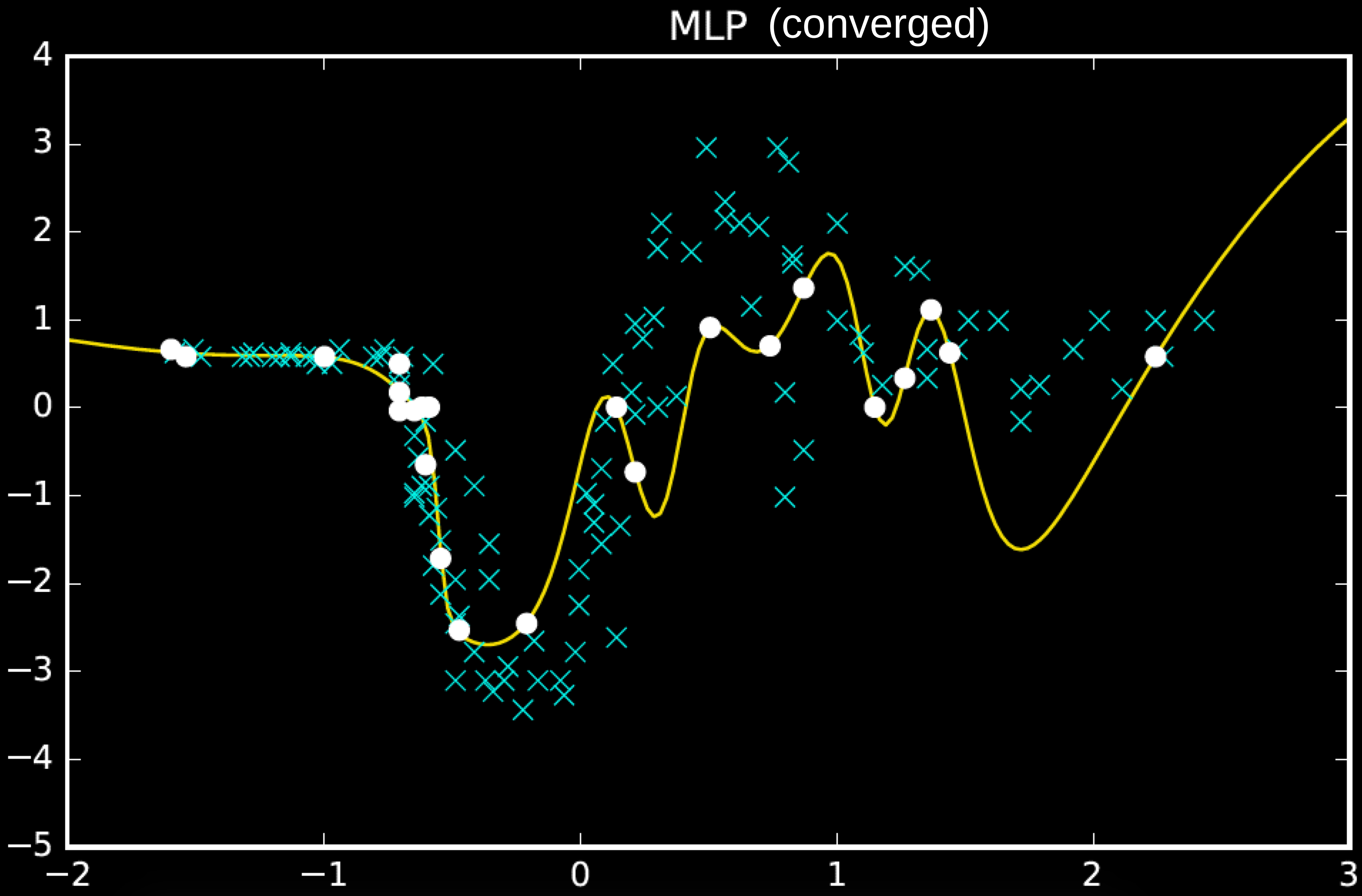
- A new approach to forming stochastic processes
- Mathematical composition:

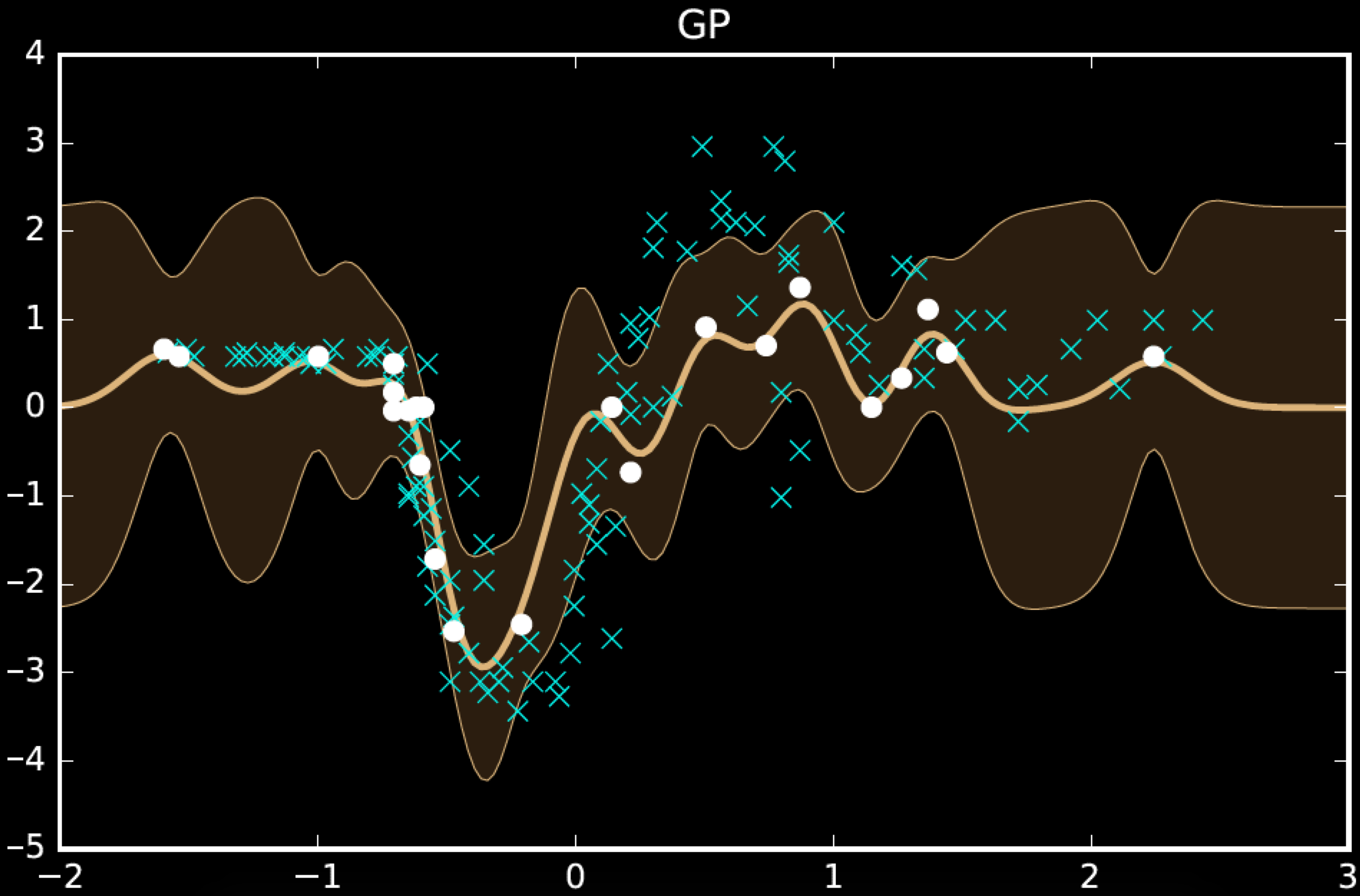
$$g(x) = f_1\left(f_2\left(f_3(x)\right)\right)$$

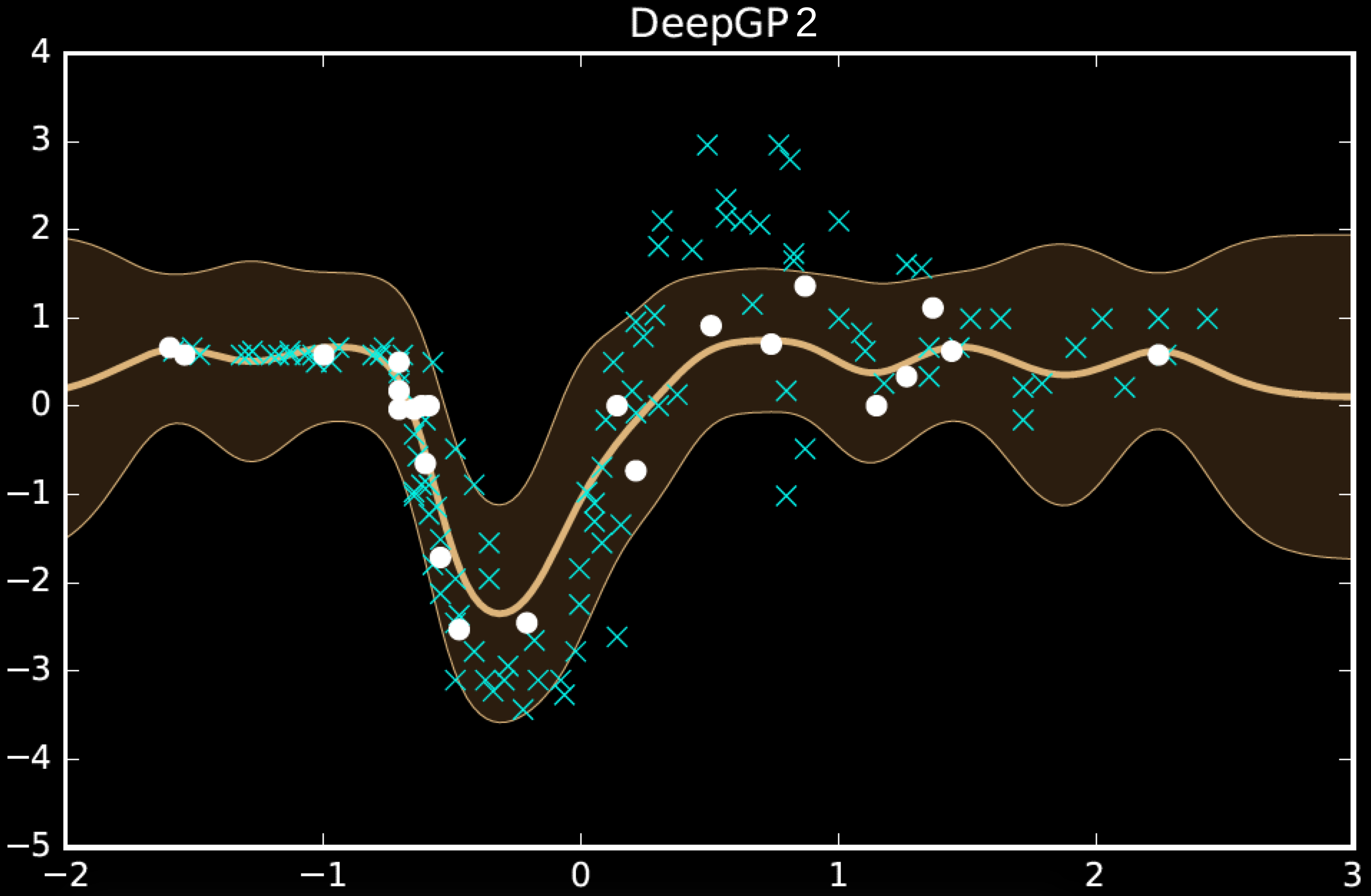
- Properties of resulting process highly non-Gaussian
- Allows for hierarchical structured form of model.

MLP (200 iterations)

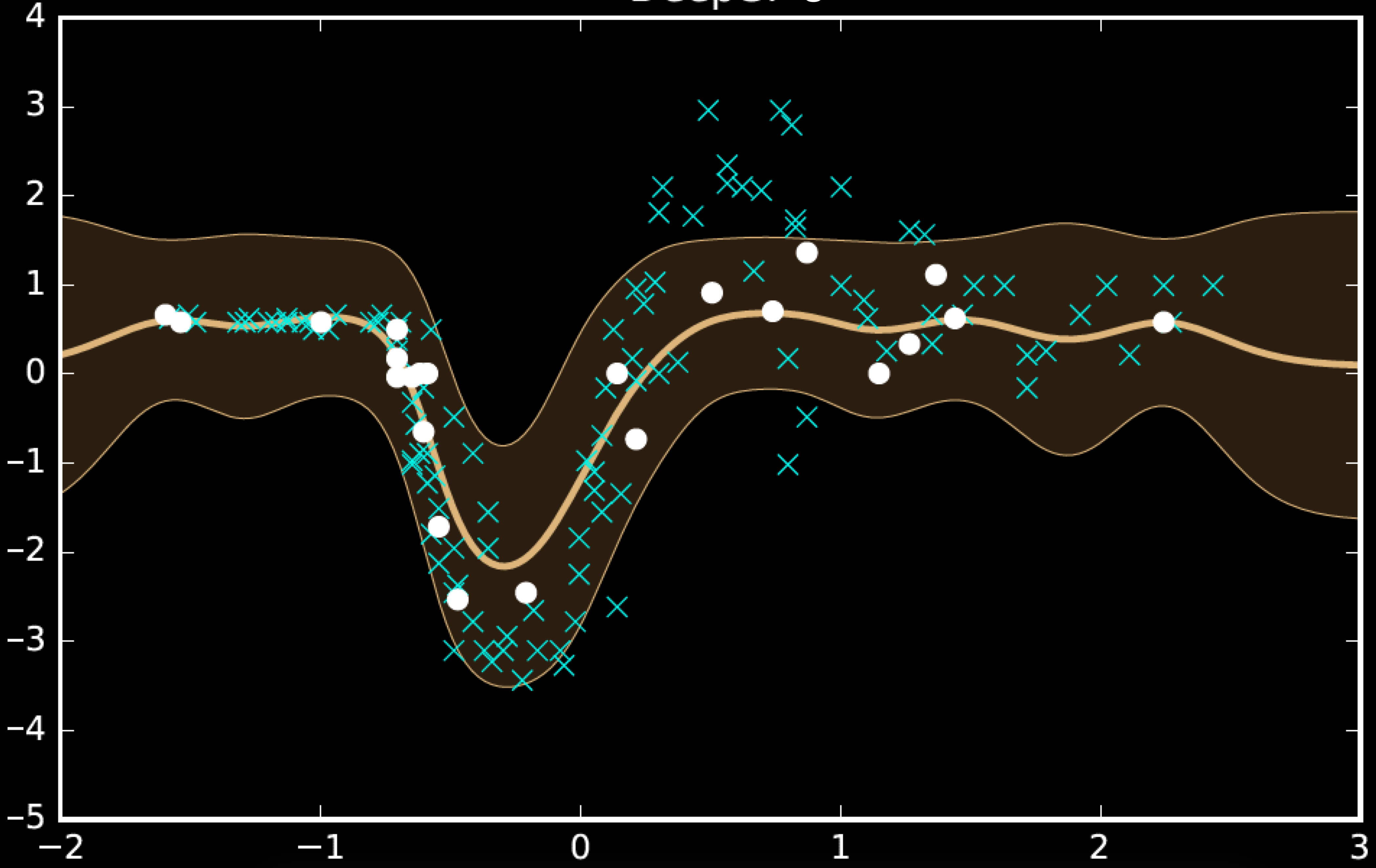



MLP (converged)
4
3
2
1
0
−1
−2
−3
−4
−5
−2
−1
0
1
2
3


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


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

DeepGP 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

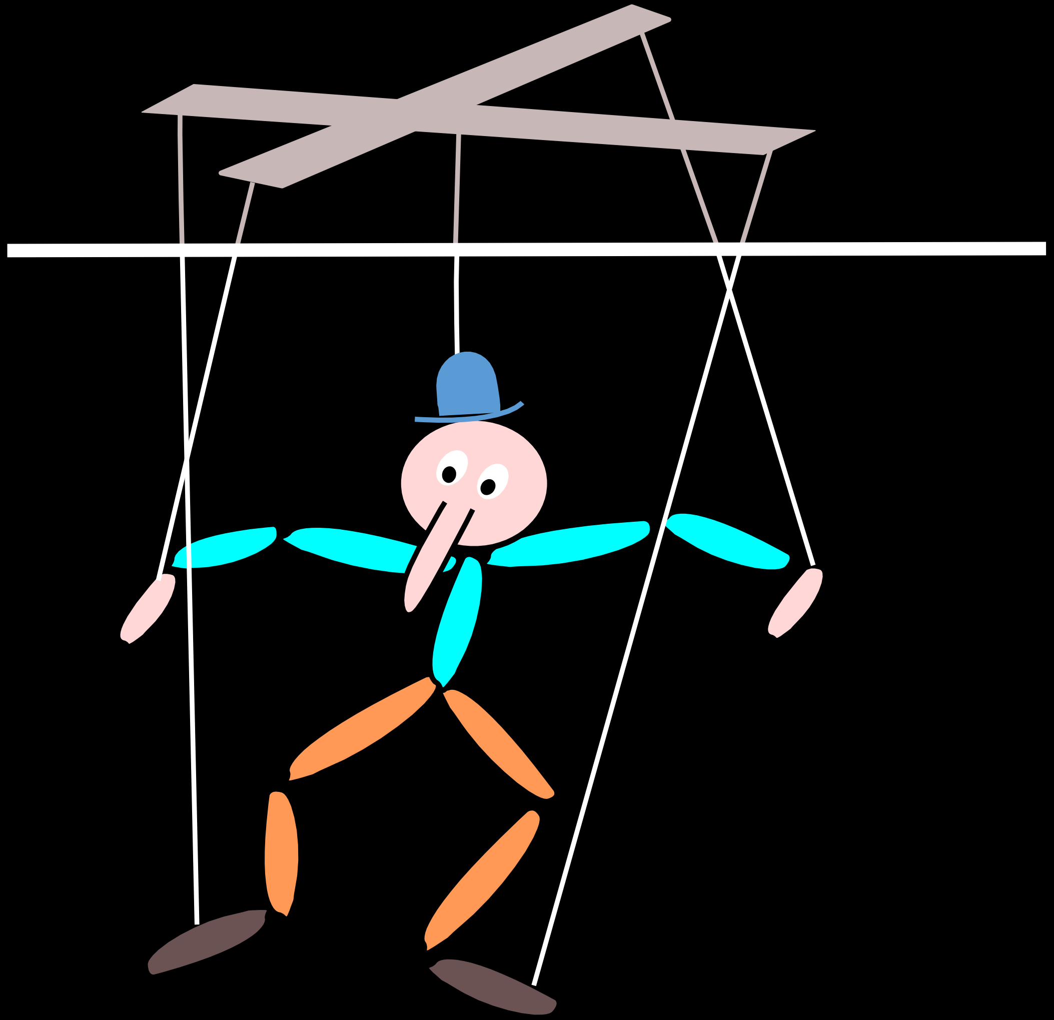
One hundred hidden nodes, one hundred inducing points

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

Classical Latent Variables

x



y

Classical Treatment

- Assume *a priori* that

$$\mathbf{x} \sim N(0, \mathbf{I})$$

- Relate linearly to $\mathbf{y}$

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \epsilon$$

- Framework covers many classical models PCA, Factor Analysis, ICA

Classical Treatment

- Assume *a priori* that

$$\mathbf{x} \sim N(0, \mathbf{I})$$

- Relate to $\mathbf{y}$ using neural net

$$\mathbf{y} = f(\mathbf{x}; \mathbf{u}, \mathbf{V}) + \epsilon$$

- Optimise over $\mathbf{u}, \mathbf{V}$

David applied importance sampling

MATLAB Demo

- demo_2016_05_03_iclr.m

New Treatment

- Assume *a priori* that

$$f(\mathbf{x}) \sim N(0, \mathbf{K})$$

- Relate to $\mathbf{y}$ using neural net

$$\mathbf{y} = f(\mathbf{x}) + \epsilon$$

- Optimise over $\mathbf{x}$

Originally inspired by density nets

MATLAB Demo

- demo_2016_05_03_iclr.m

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

expected
log likelihood

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

model remains linear in $\mathbf{u}$

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

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

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

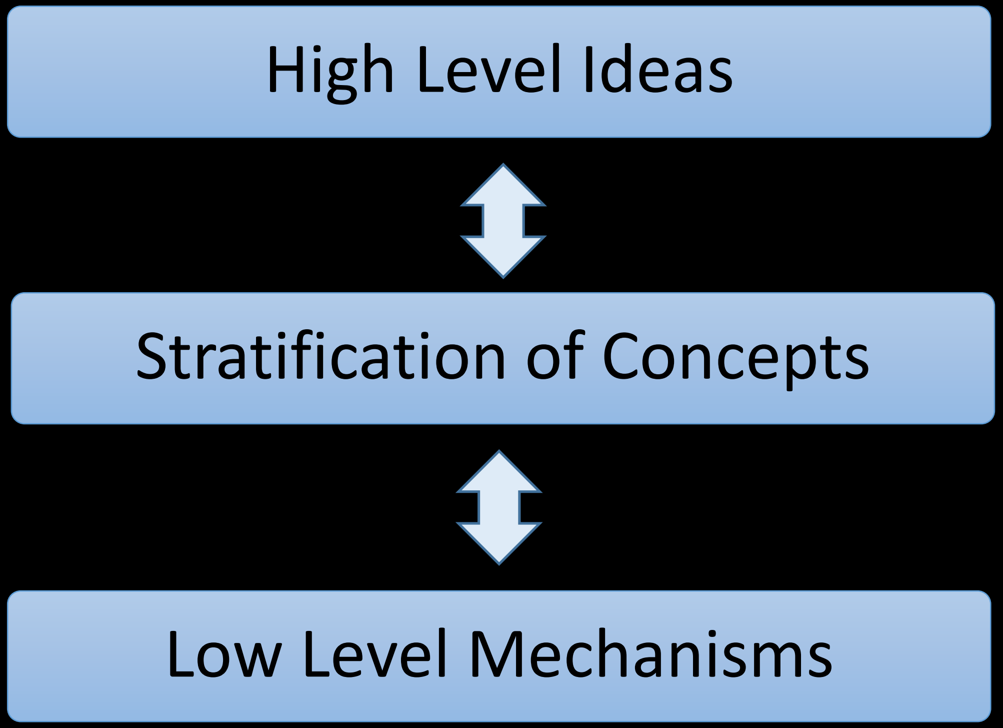
model is not linear in $\mathbf{x}$

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

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

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

Use Abstraction for Complex Systems

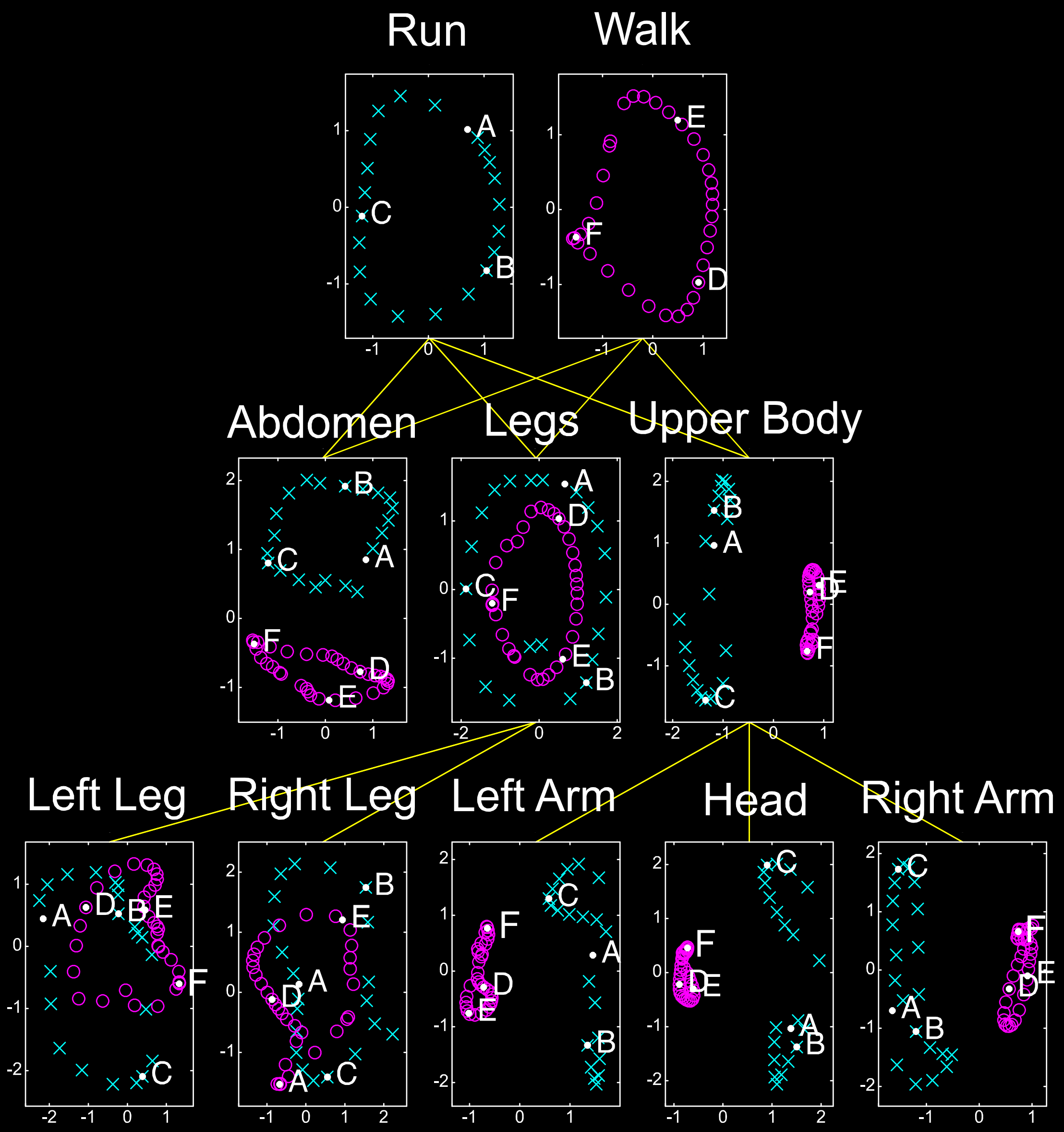


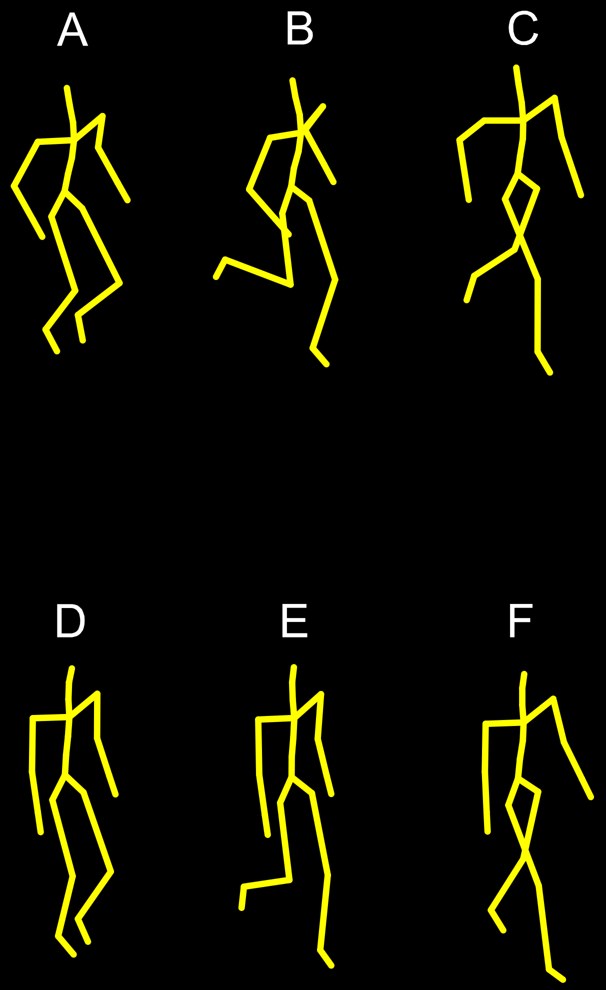
High Level Ideas

Stratification of Concepts

Low Level Mechanisms

Example: Motion Capture Modelling

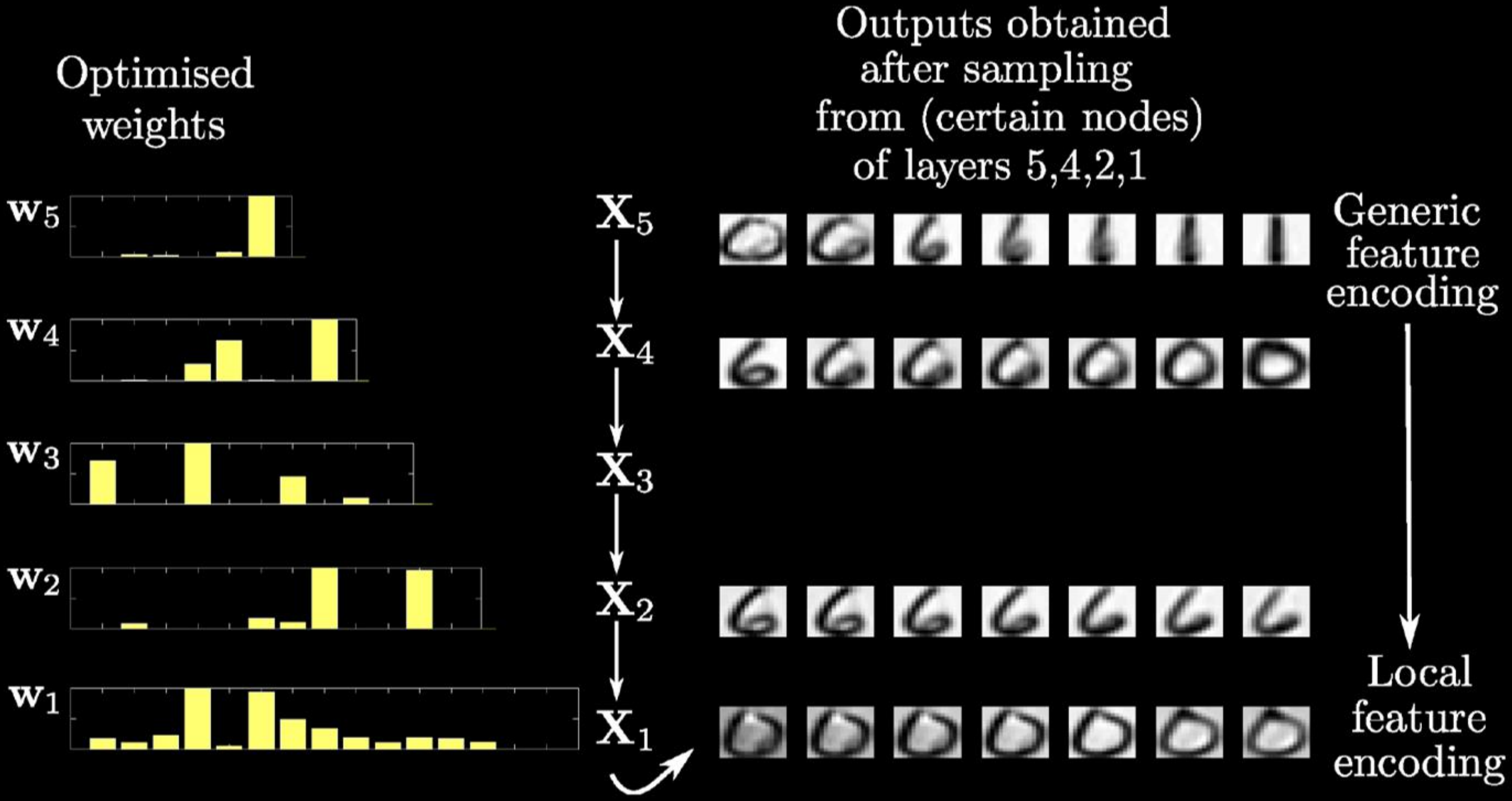




MATLAB Demo

- demo_2016_05_03_iclr.m

Modelling Digits

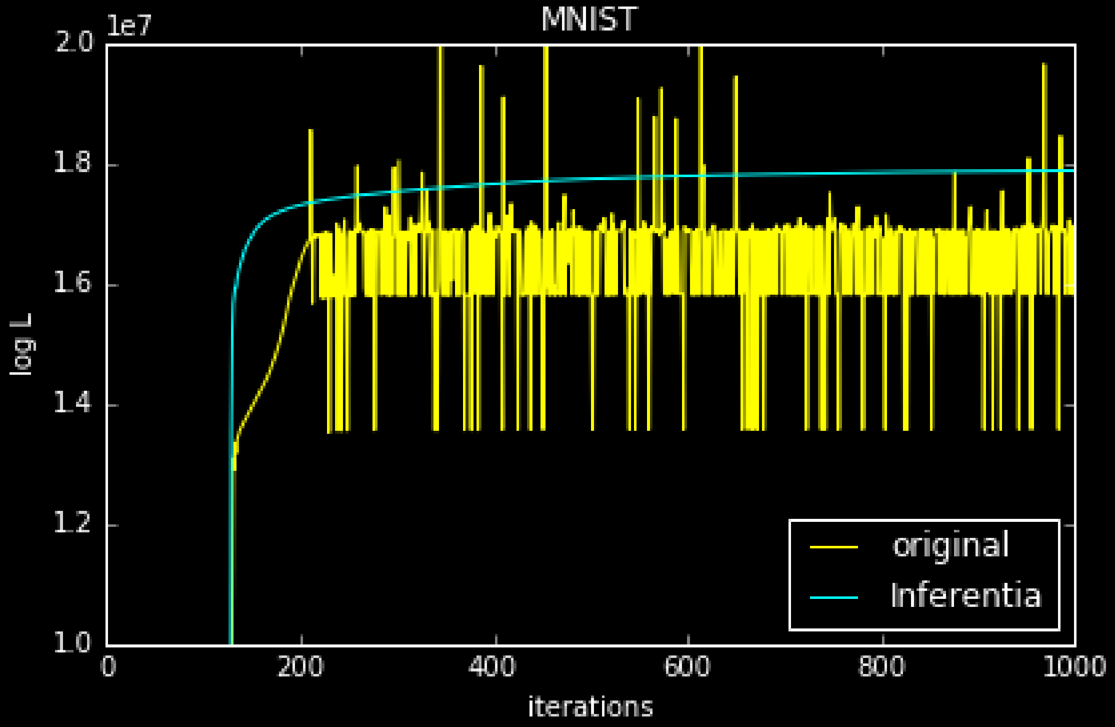


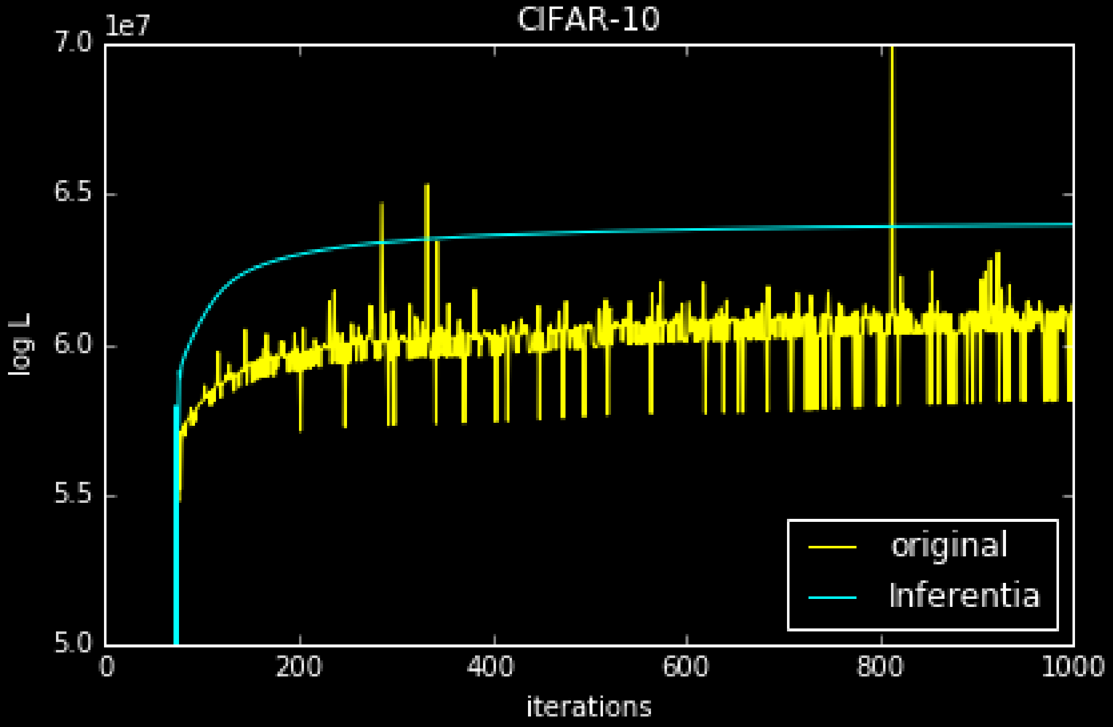
MATLAB Demo

- demo_2016_05_03_iclr.m

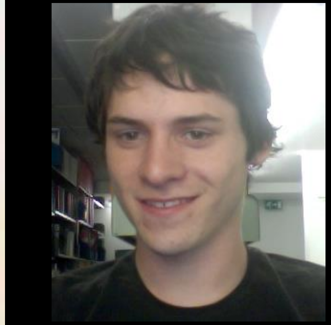
Inferentia

Numerical Issues

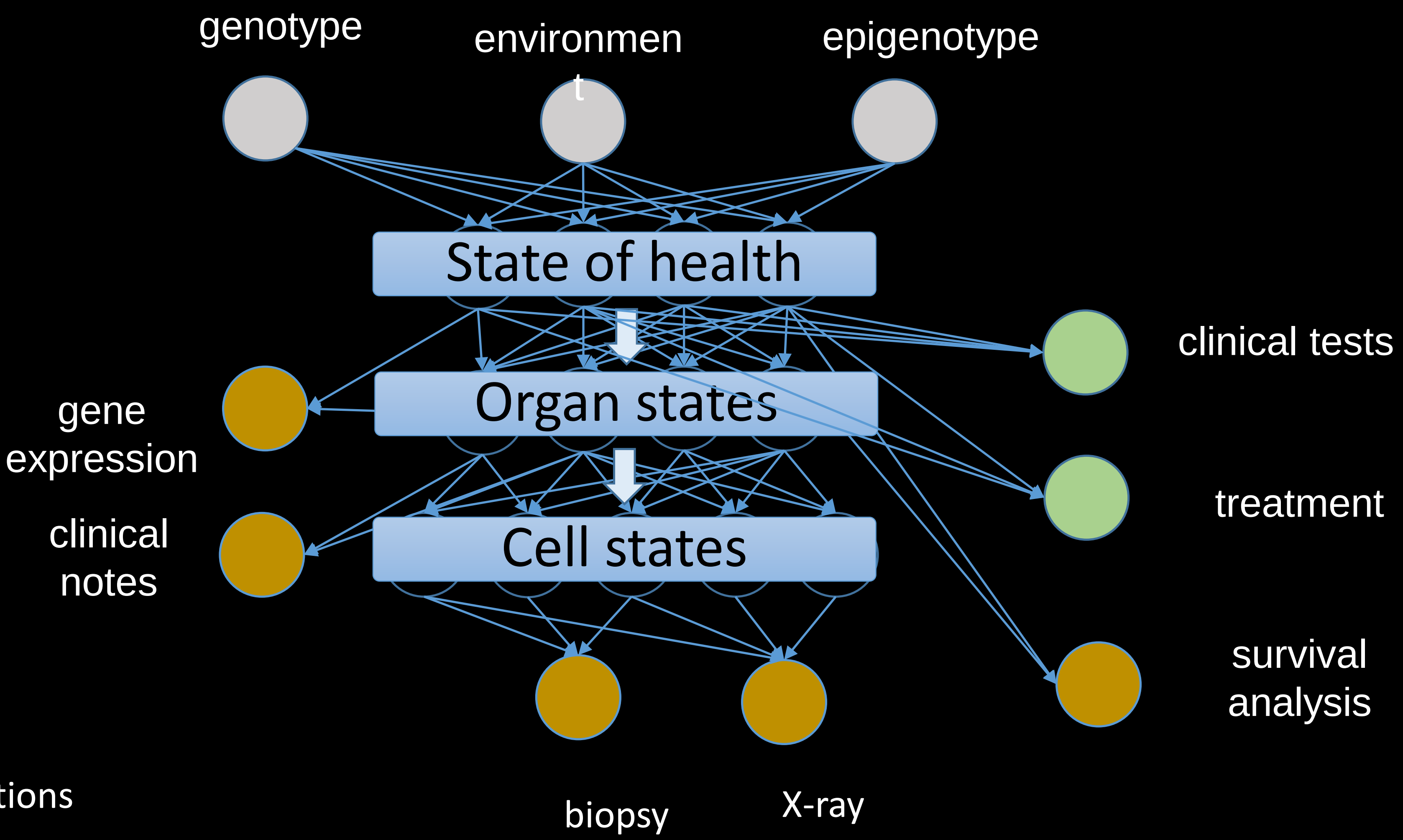




Health



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



PLoS Comp Bio, Nature Communications

To Find Out More

- Gaussian Process Summer School
  - 12th-15th September 2016 in Sheffield http://gpss.cc/
- Posters at ICLR:
  - Recurrent Gaussian Processes
  - Variationally Auto-Encoded Deep Gaussian Processes
- Python software for GPs (GPy)
  - https://github.com/SheffieldML/GPy/

David's "Gaussian Process Basics" talk

Thank you

Neil Lawrence
http://inverseprobability.com
@lawrennd