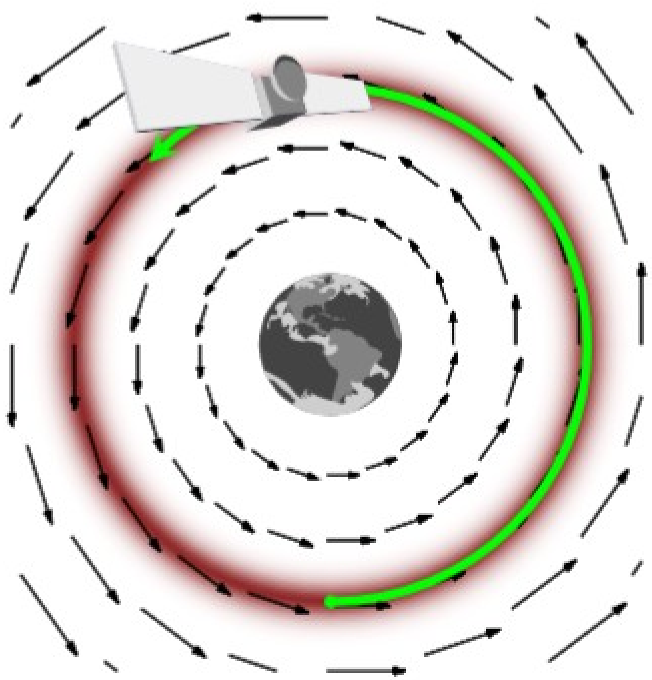
University of Amsterdam

Machine learning for strong lensing image analysis

Gradients



Gravity

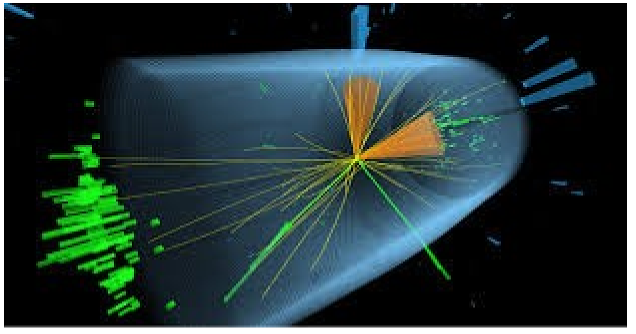
Christoph Weniger

1909.xxxxx: Marco Chianese, Adam Coogan, Paul Hofma, Sydney Otten, CW

25 Sep 2015, PALS meeting Paris

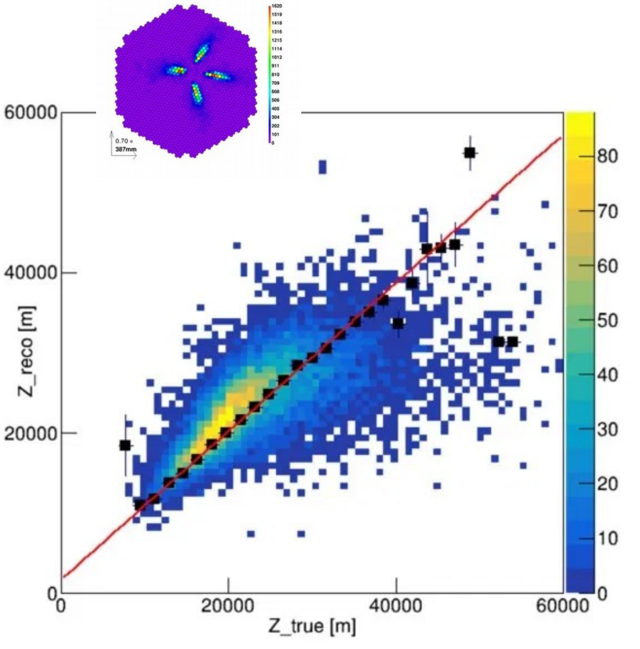
Deep learning in HEP & Astronomy

LHC event reconstruction



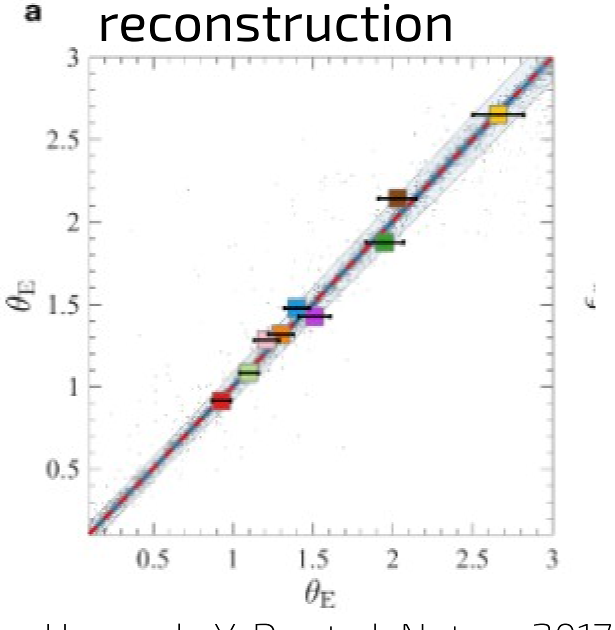
https://arxiv.org/abs/1806.11484

CTA event reconstruction



http://arxiv.org/abs/1810.00592

Gravitational lensing reconstruction



Hezaveh, Y. D., et al. Nature 2017.

Deep learning (in particular supervised learning) techniques are becoming increasingly more important for HEP & Astronomy.

→ Faster and more precise inference

Deep neural networks

Deep neural networks are extremely flexible function approximators

Flexible function

Input: $\vec{x}$ $$f_\phi(\vec{x}) = \vec{y}$$ Output : $\vec{y}$

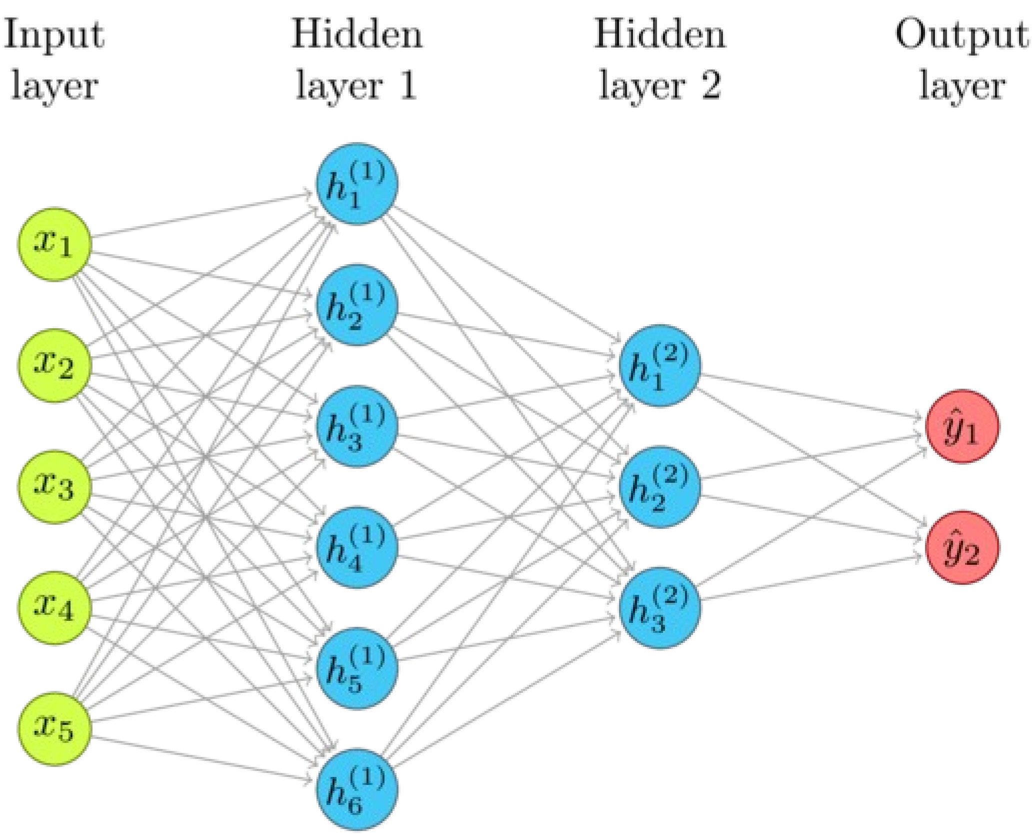
ϕ : Internal network parameters/weights

Examples

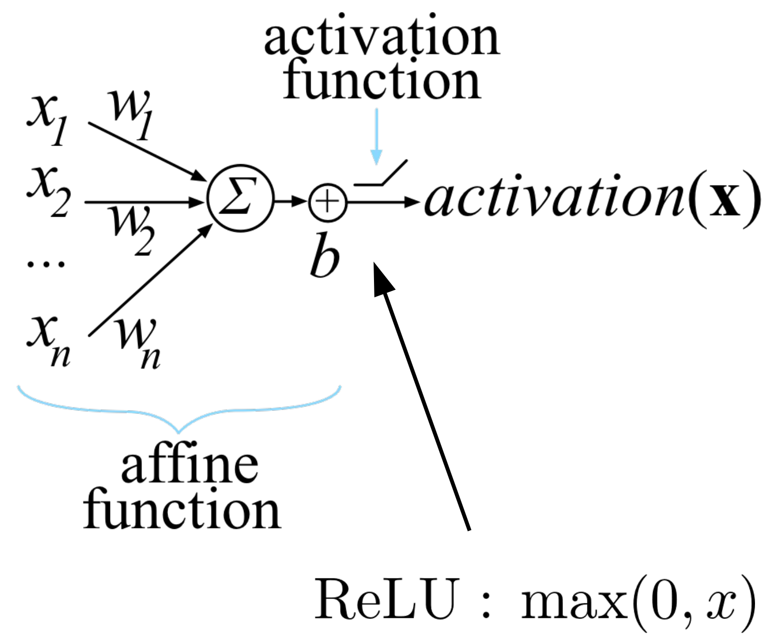
Image	Classification →	Discrete labels
Image	Regression →	Continuous parameters
Parameters	Generative model →	Image
Image	e.g. denoising →	Image

Deep neural networks

Typical feed-forward network with multiple hidden layers

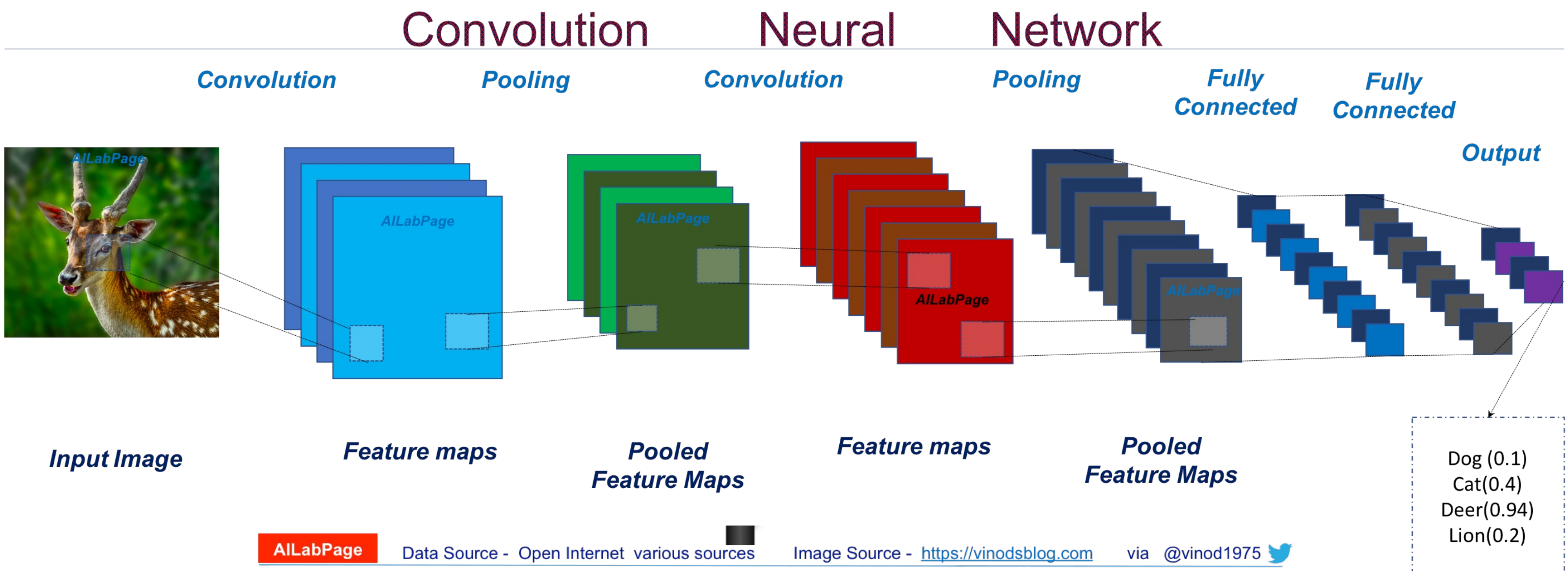


Layer connection: Stacking affine and activation function

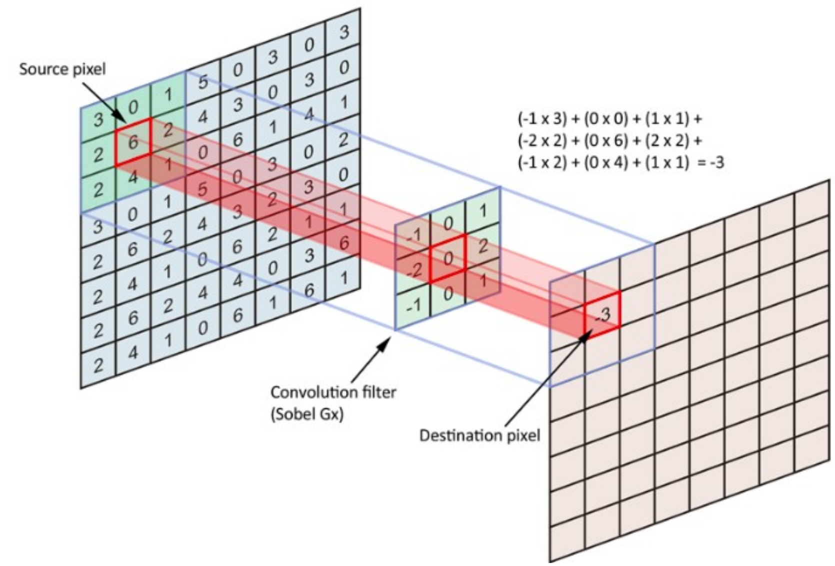


Network weights and biases $\phi = \vec{w}, \vec{b}$

Example: Image classification network



Convolutional layer



Pooling layer

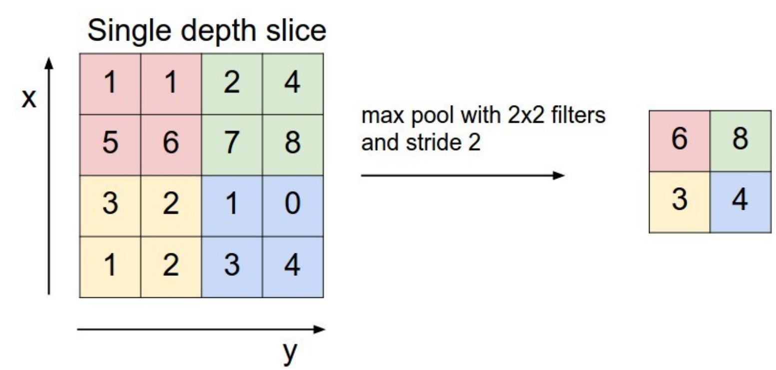


Image credit: https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2, http://cs231n.github.io/convolutional-networks/

Training with gradient decent

1. Define loss function
   (here square error / chi-square)

$$\text{loss} = \sum_{\text{batch}} (f_\phi(\vec{x}_i) - \vec{y}_i)^2$$

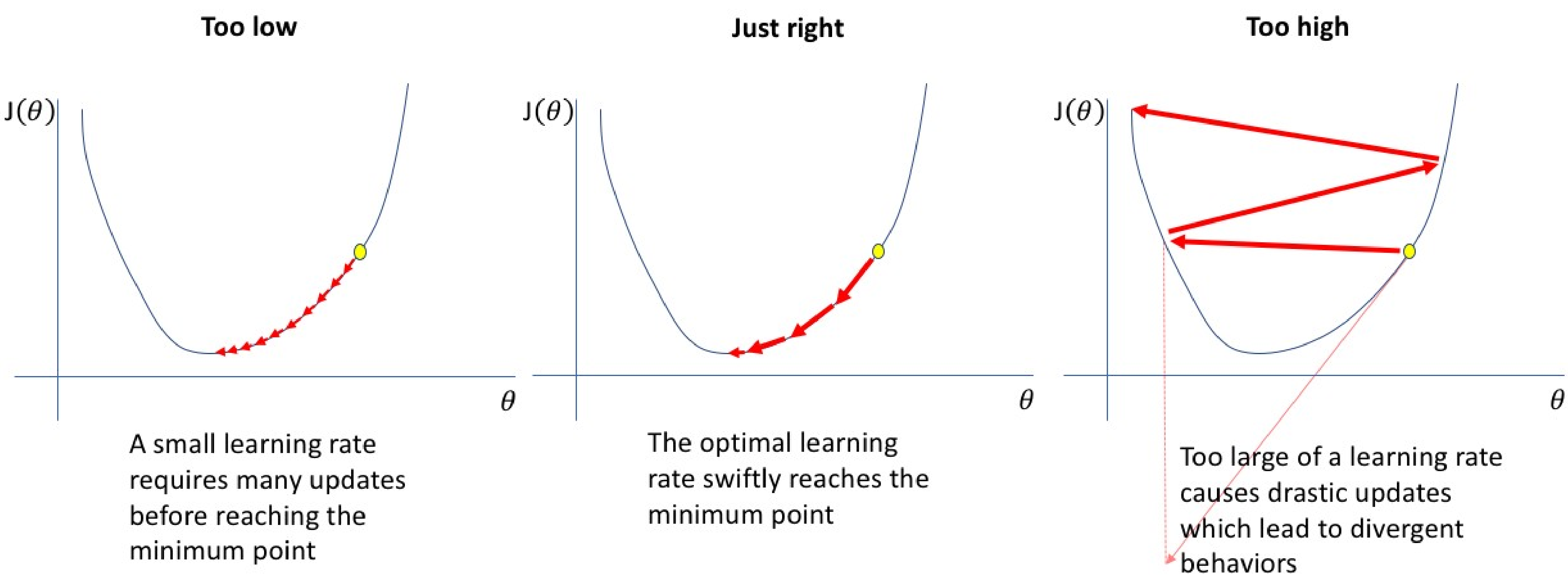
2. back-propagation

$$\text{grad} = \frac{\partial \text{loss}}{\partial \phi}$$

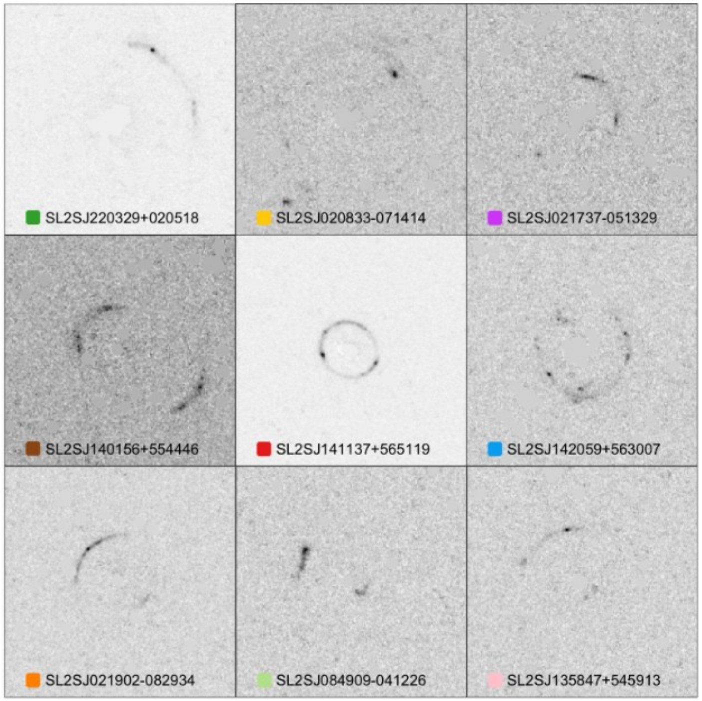
3. parameter update

$$\phi \to \phi - \text{LR} \cdot \text{grad}$$

Stop when it works well on separate *test data*

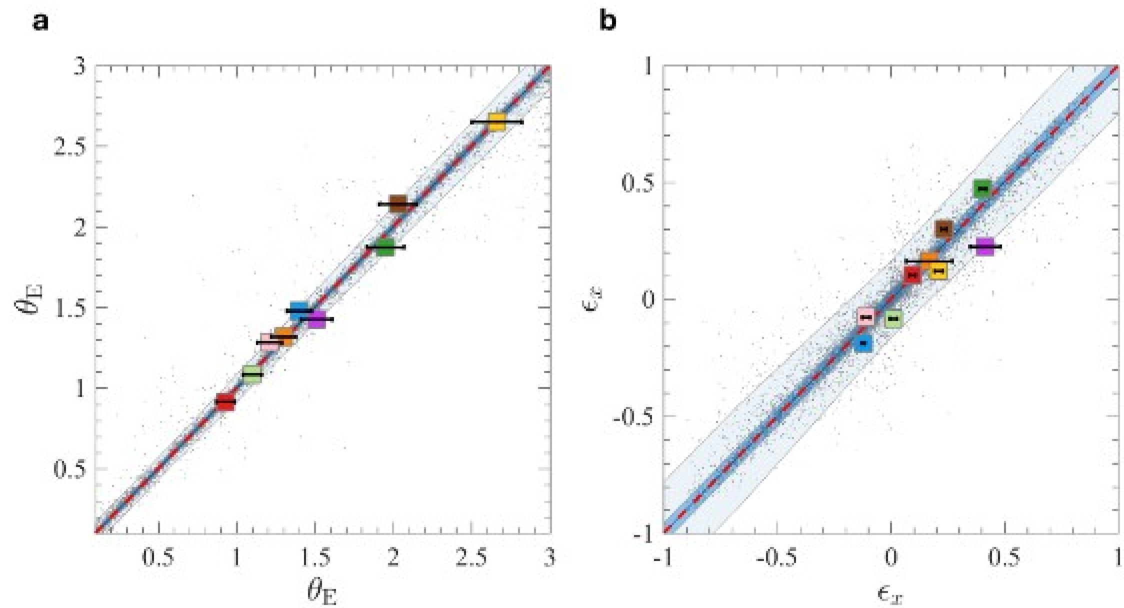


Example: Gravitational Lensing



Gravitational lensing data (HST)

Reconstructed parameters



- Amazing speed up: 1 day → 0.01 s
- Somewhat less accurate
- Challenge: Subhalos will show up as O(1%) perturbations of best-fit model, unclear if models can be trained for that as well?

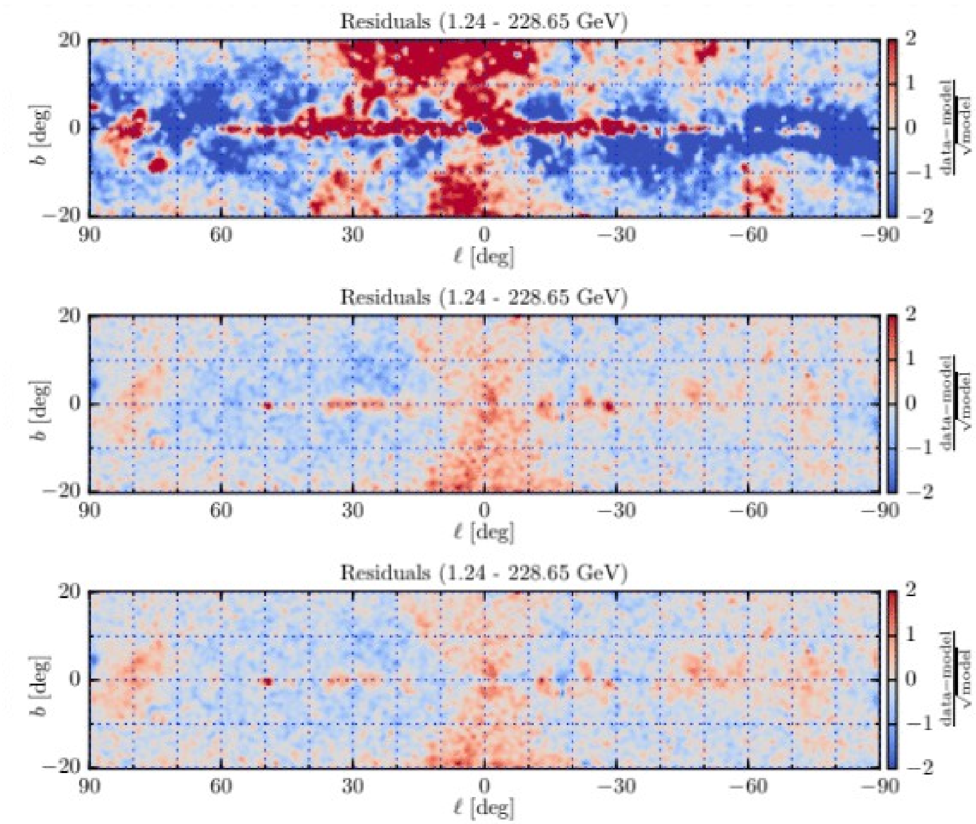
Hezaveh, Y. D., Levasseur, L. P. & Marshall, P. J. Fast Automated Analysis of Strong Gravitational Lenses with Convolutional Neural Networks, Nature 2017.

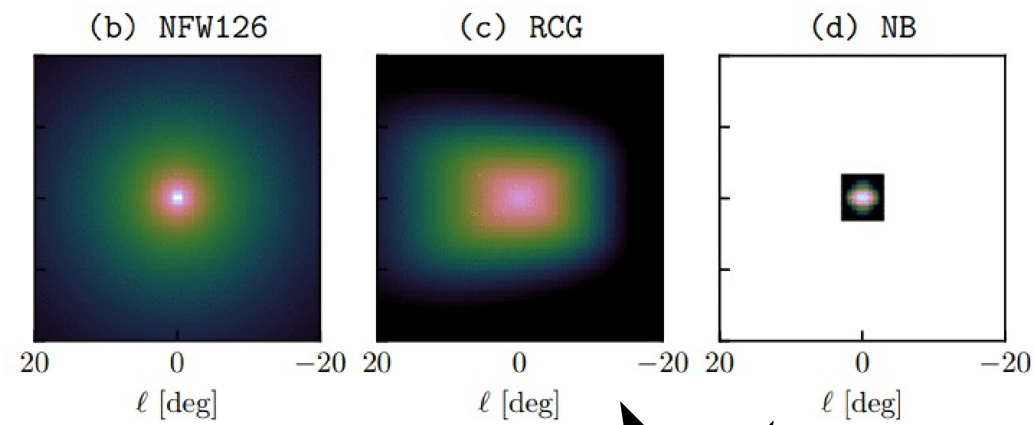
What else are gradients good for?

Source: Deep Ideas

1) High-dimensional optimization

Some exemplary residuals of gamma-ray emission along the Galactic disk, accounted for with 1e5 of nuiscance parameters. We use L-BFGS-B algorithm for optimization.





Bulge preferred over DM

Bartels+, Nature Astronomy 2018

Very-high dimensional models (for e.g. for gamma-ray emission in the Galactic disk) would be extremely hard to optimize without gradient information.

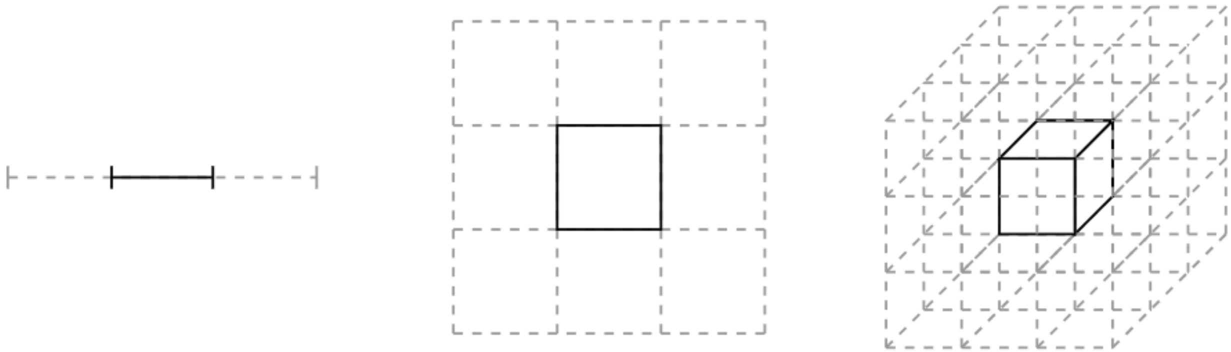
2) High-dimensional sampling

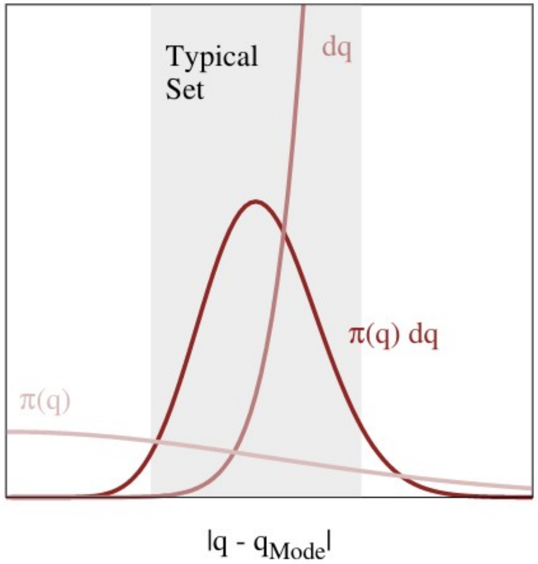
Typical goal of **Bayesian inference** is to evaluate expectation values of the form

$$\mathbb{E}_\pi[f] = \int_Q \mathrm{d}q \, \pi(q) \, f(q)$$

Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo (1701.02434).

The course of dimensionality

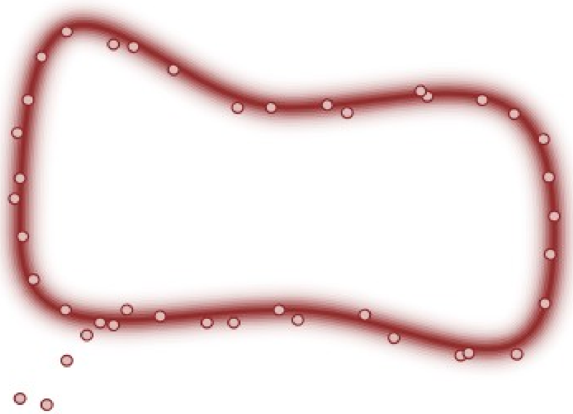




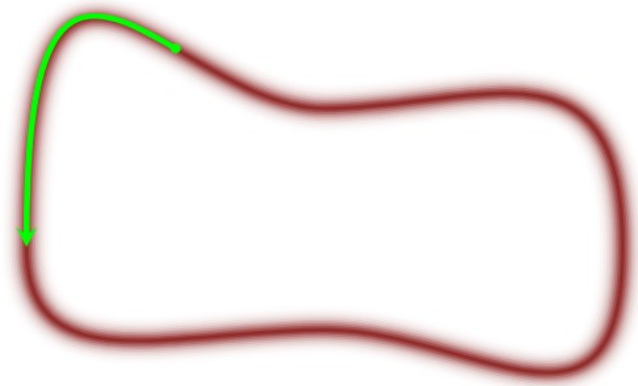
The integral is dominated by points in the "**typical set**", which is *not* conincident with the mode if dimensionality is large

2) Hamiltonian Monte Carlo

Goal is to explore the "typical set", using Monte Carlo techniques.

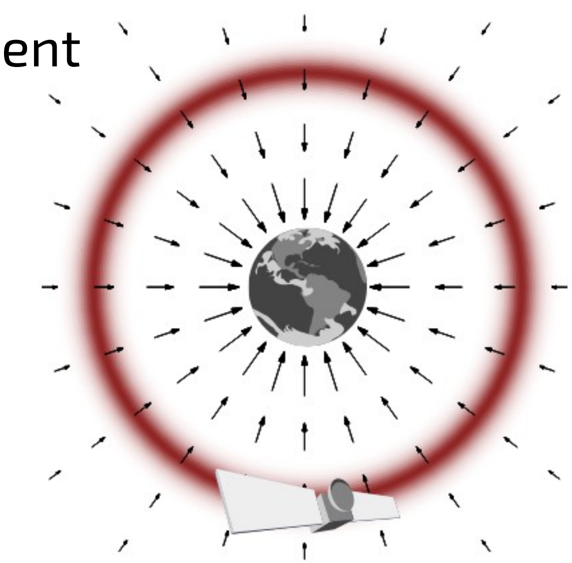


Knowing the direction of the typical set would greatly help.



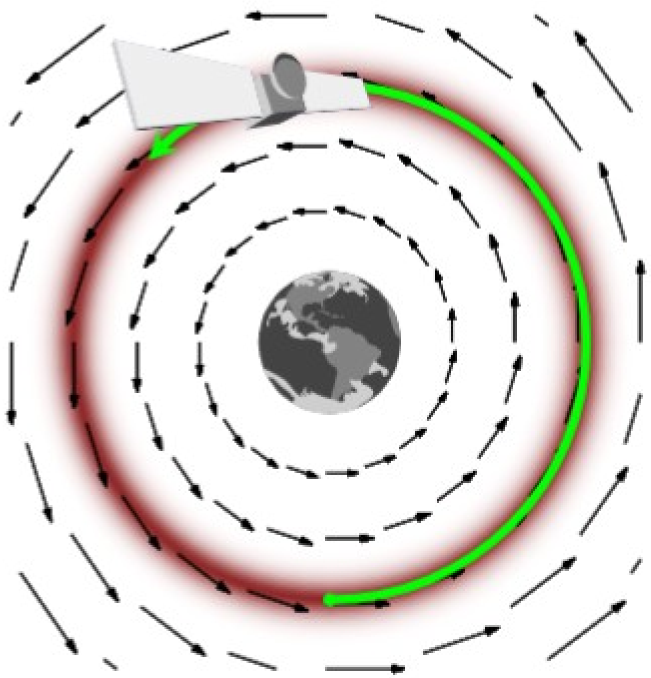
However, simply following the gradient would just lead to the mode

$$V(q) = -\ln \pi(q)$$



2) Hamiltonian Monte Carlo

Idea: Introduce momentum, *p*, such that MC is kept in the "orbit" of highest density mass (in the 'typical set').



Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo (1701.02434).

New target function

$$\pi(q,p) = e^{-H(q,p)}$$

$$\begin{aligned} H(q,p) &= -\log \pi(p \mid q) - \log \pi(q) \\ &\equiv \quad K(p,q) \quad + \ V(q) \,. \end{aligned}$$

Sampling

- Pick initial point
- follow Hamiltonian dynamics

$$\frac{\mathrm{d}q}{\mathrm{d}t} = +\frac{\partial H}{\partial p} = \frac{\partial K}{\partial p}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial H}{\partial q} = -\frac{\partial K}{\partial q} - \frac{\partial V}{\partial q}$$

- resample momentum from canonical distribution, preserve position

3) Variational Inference

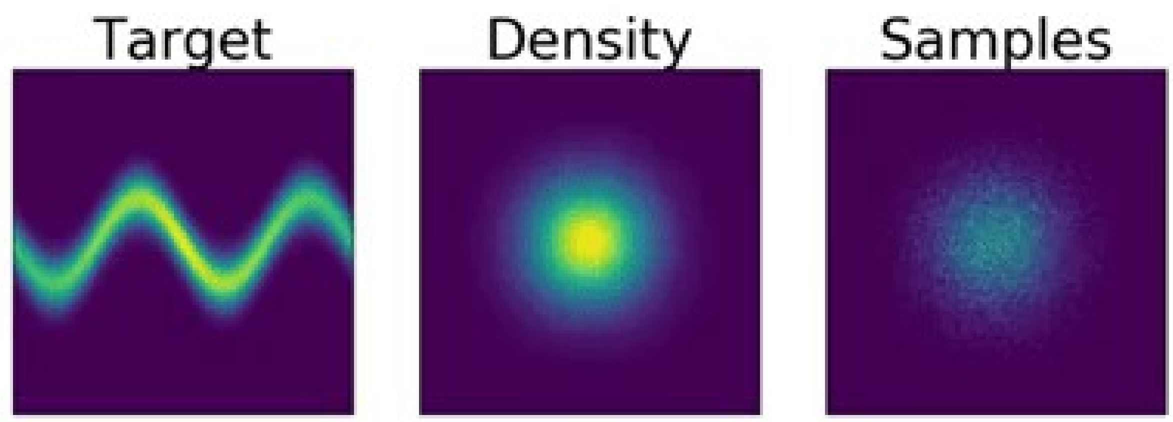
Kullback-Leibler divergence:

$$\mathrm{KL}(q(z)||p(z|x)) = \int dz q(z) \ln \frac{q(z)}{p(z|x)} = \mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log p(\mathbf{z} \mid \mathbf{x})]$$

Fit parameteric model for posterior, q(z), to true posterior p(z|x).

Sampling problem → Optimization problem

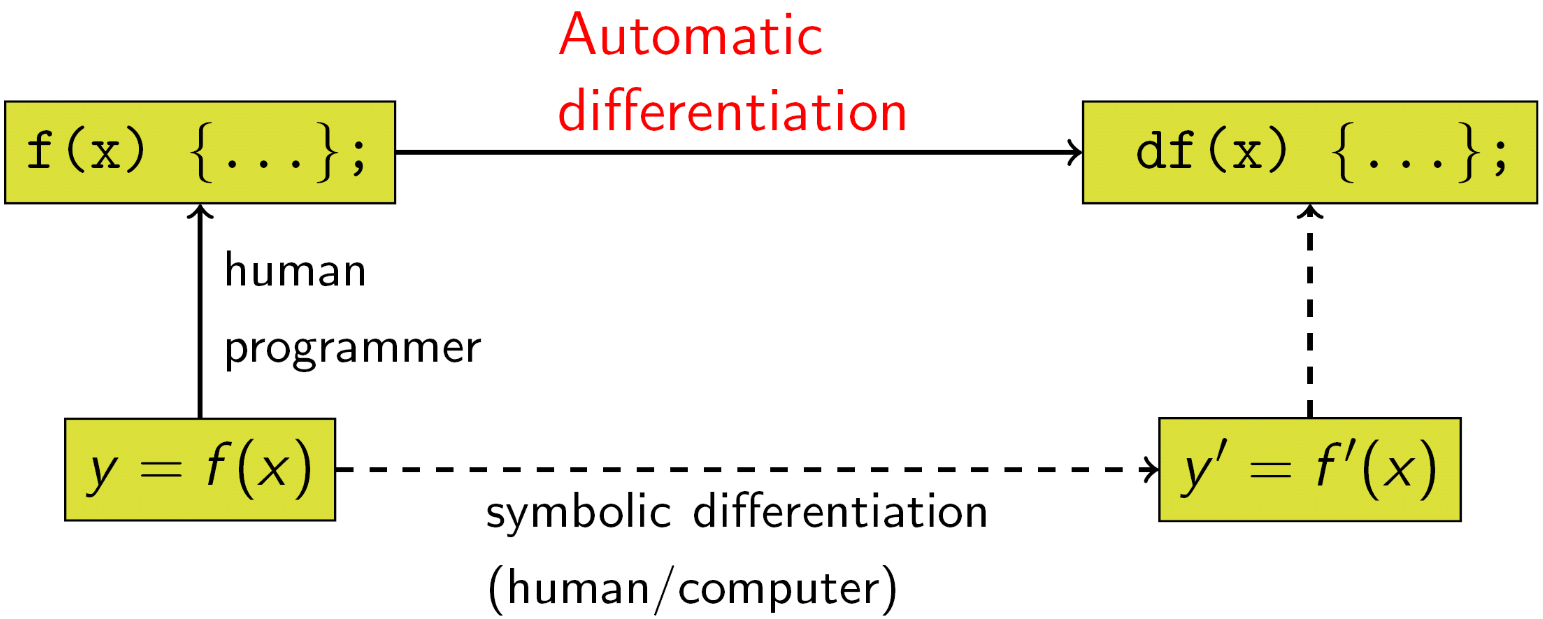
$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathscr{Q}}{\arg\min} \, \mathrm{KL}\,(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}))$$



Video credit: Jesse Bettencourt

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians (1601.00670).

Many recent auto-grad tools



Forward-mode Differentiation: $h_i \equiv \frac{\partial}{\partial x} f_i(x)$

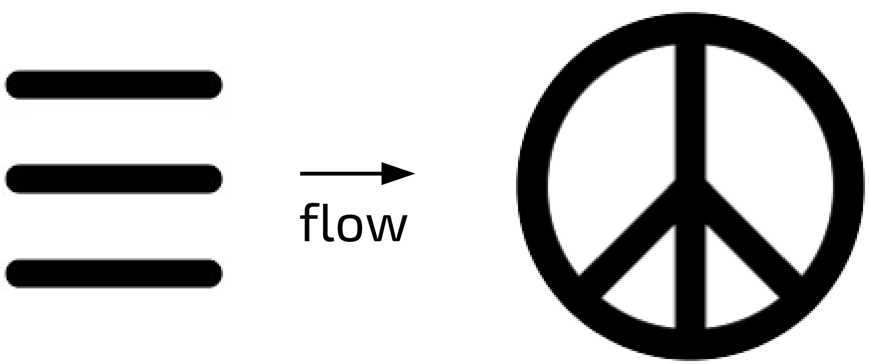
Back-propagation: $g_i \equiv \frac{\partial}{\partial x_i} f(\vec{x})$

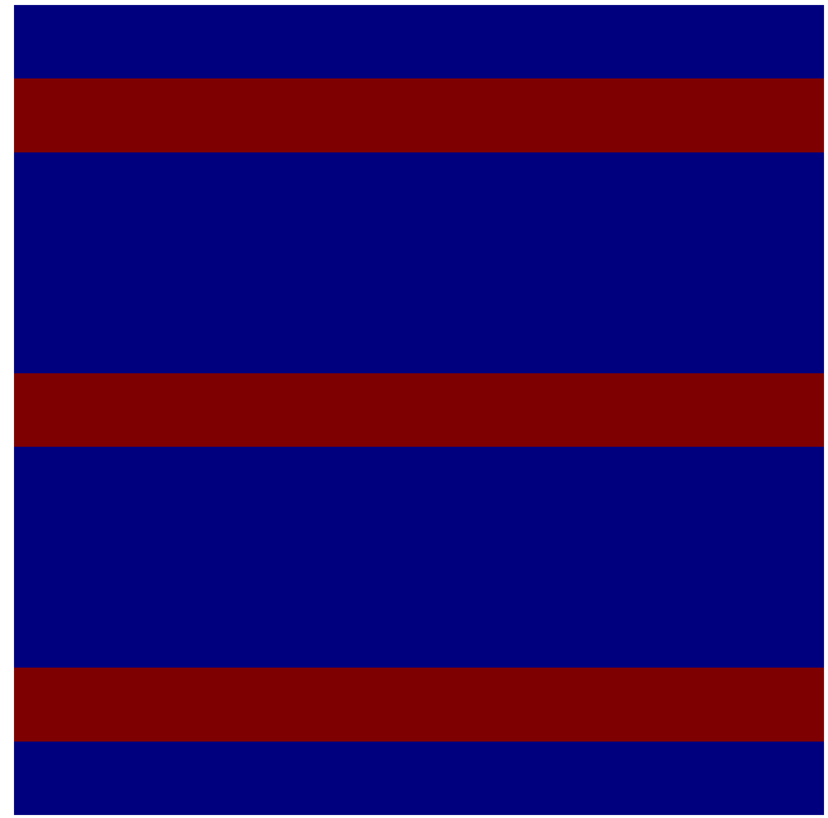
https://medium.freecodecamp.org/demystifying-gradient-descent-and-backpropagation-via-logistic-regression-based-image-classification-9b5526c2ed46

Auto-grad through Euler fluid equations

Optimization goal

Find initial velocity field that does this



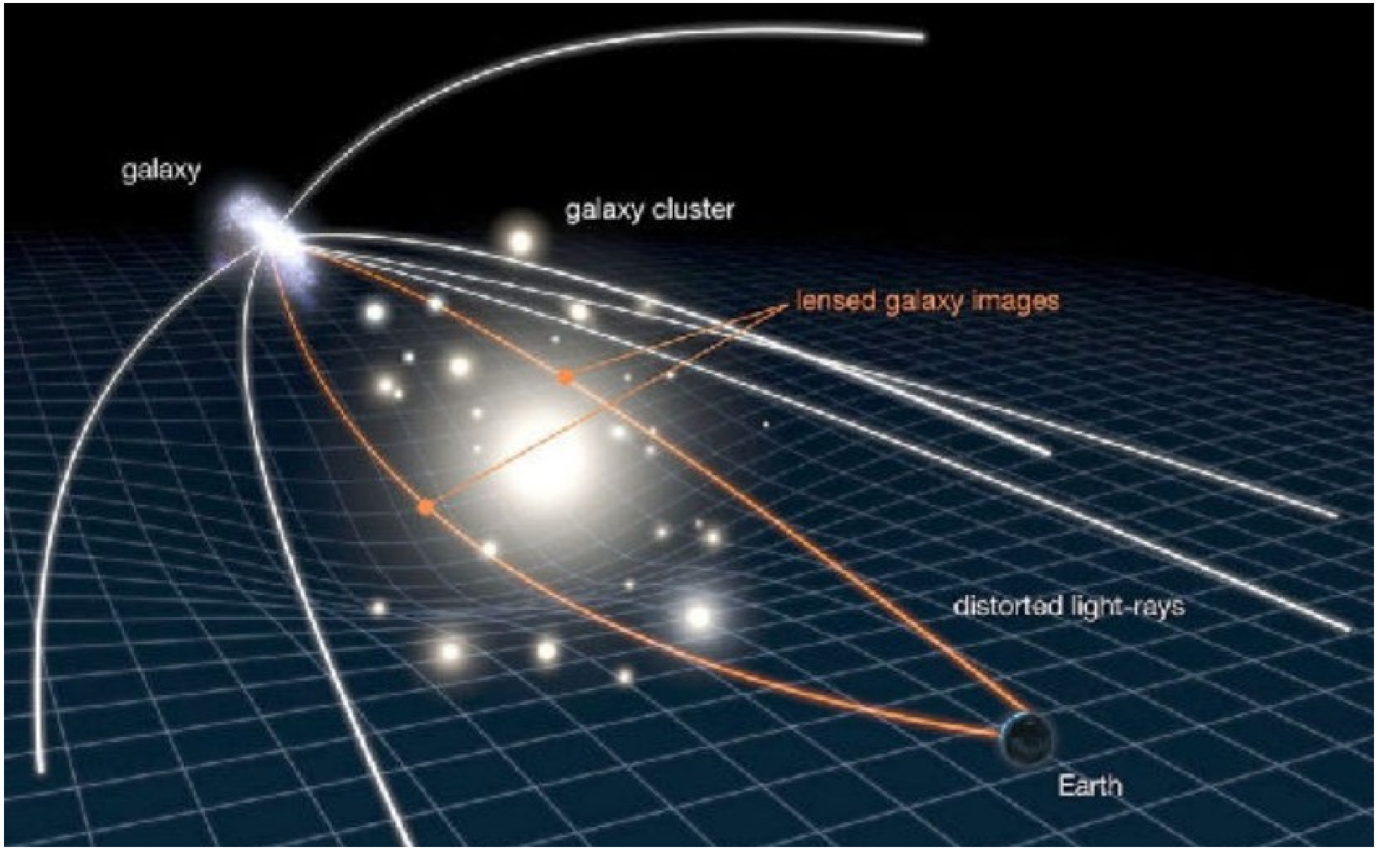


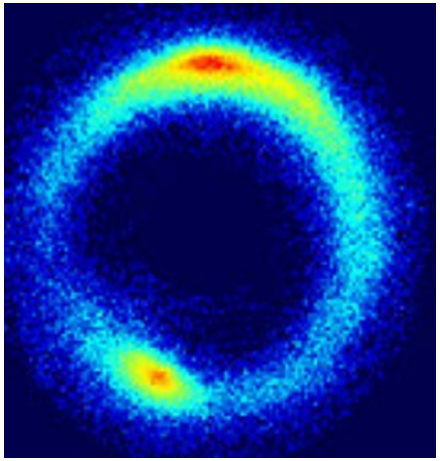
https://github.com/HIPS/autograd

Strong gravitational lensing

Strong gravitational lensing

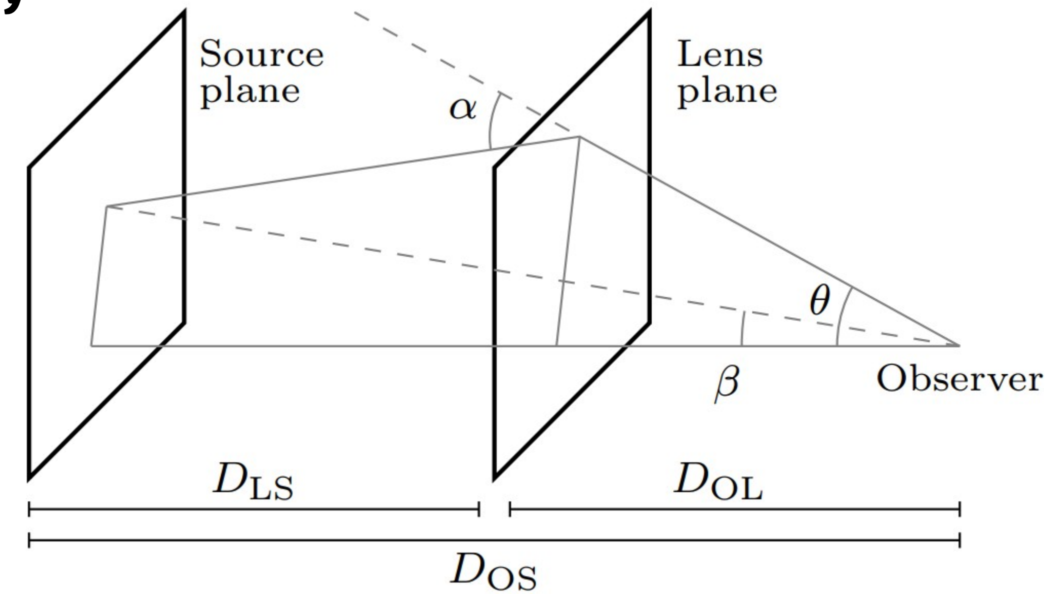
- Light from distant galaxies is deflected by DM halos long the line of sight
- Leads to multiple images, arcs, near-perfect Einstein rings
- Careful analysis of lensed images reveals information about DM halos





Strong lensing basics

Basic geometry



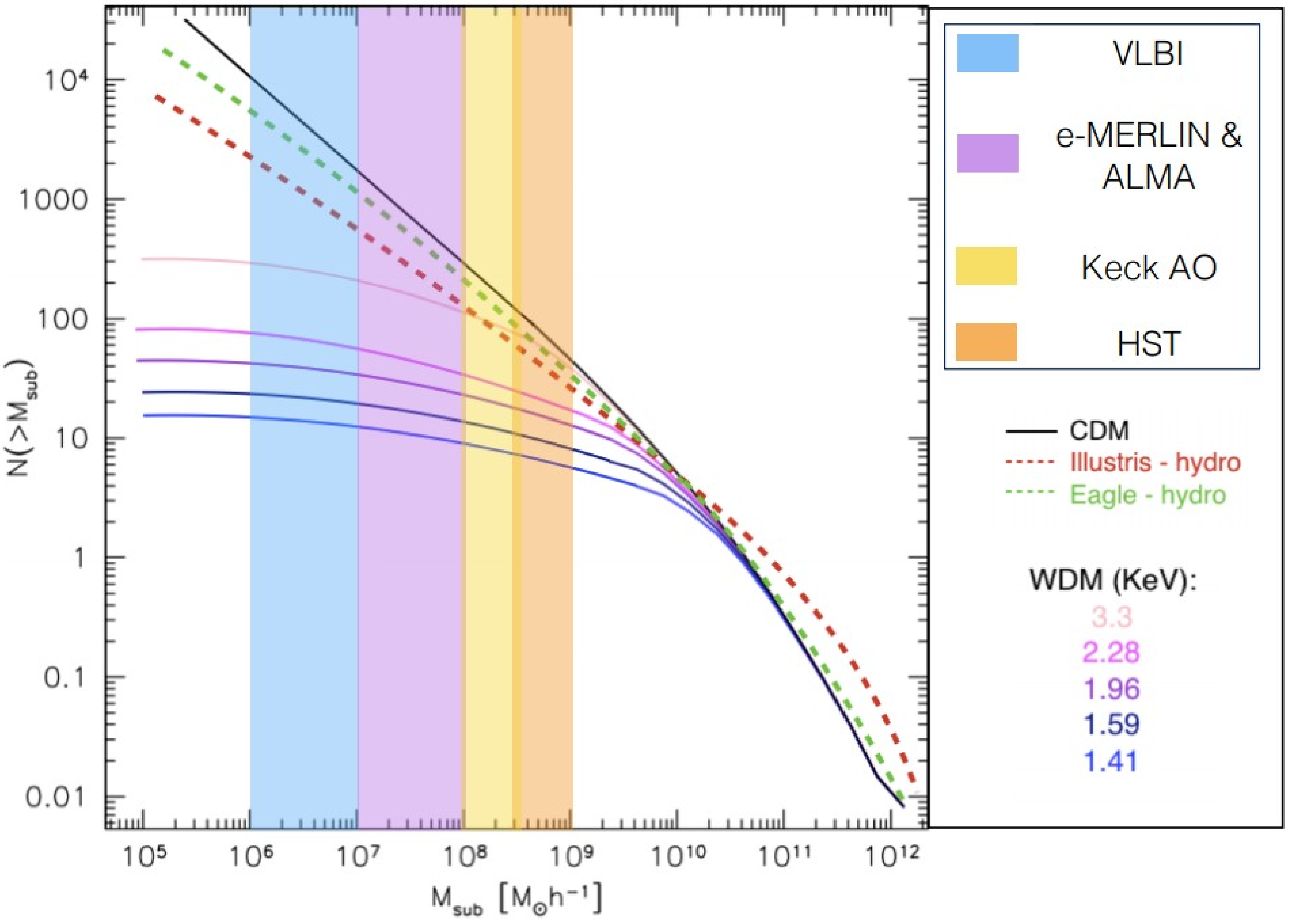
Displacement field from Poisson kernel convolution

$$\boldsymbol{\alpha} = \frac{4\,G}{c^2} \frac{D_{\mathrm{OL}} D_{\mathrm{LS}}}{D_{\mathrm{OS}}} \int \Sigma(\boldsymbol{\theta}') \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2} d^2\theta'$$

Lensed image

$$\mathcal{I}_{\mathrm{lens}}(\boldsymbol{\theta}) = \mathcal{I}_{\mathrm{src}}(\boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta}))$$

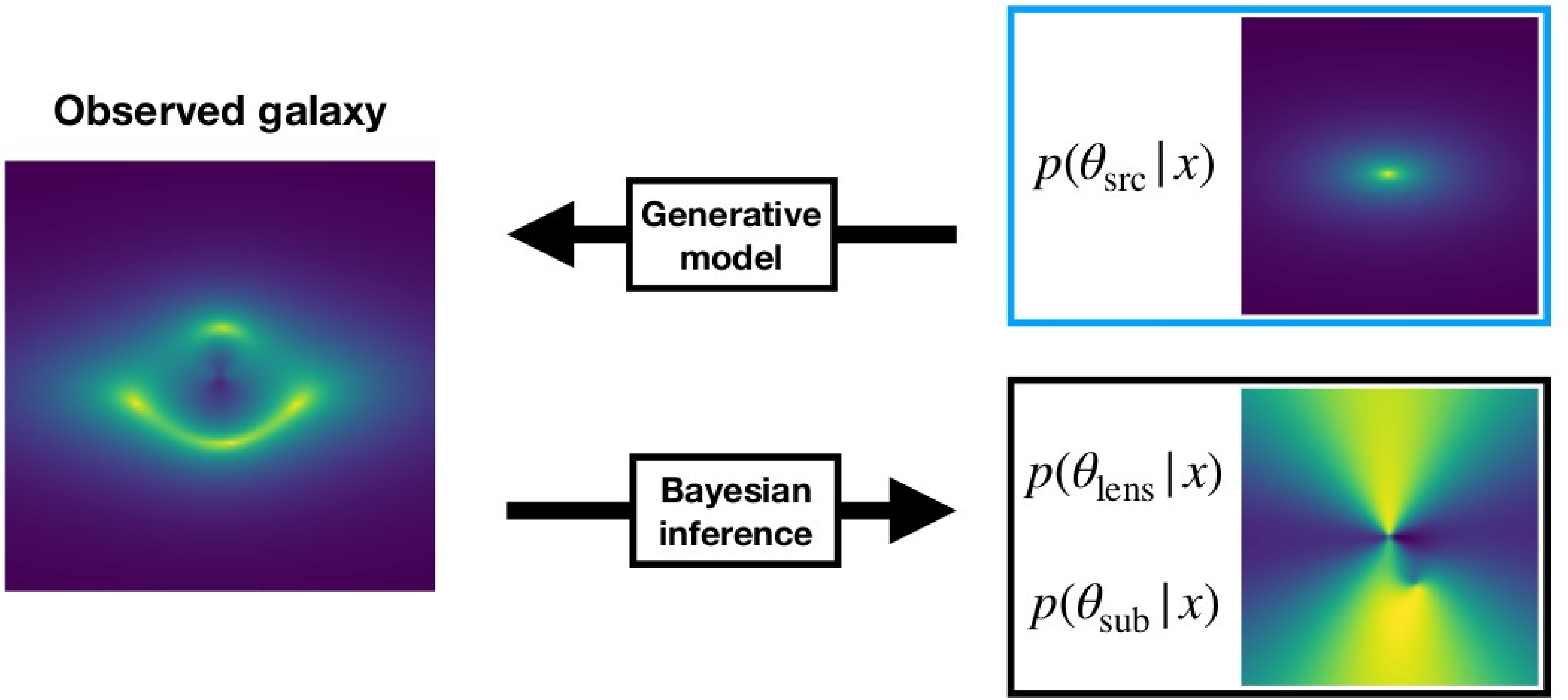
Probe for DM temperature & mass



The cut-off in the mass function is directly related to the model for dark matter.

Slide credit: John McKean (ASTRON) 2018

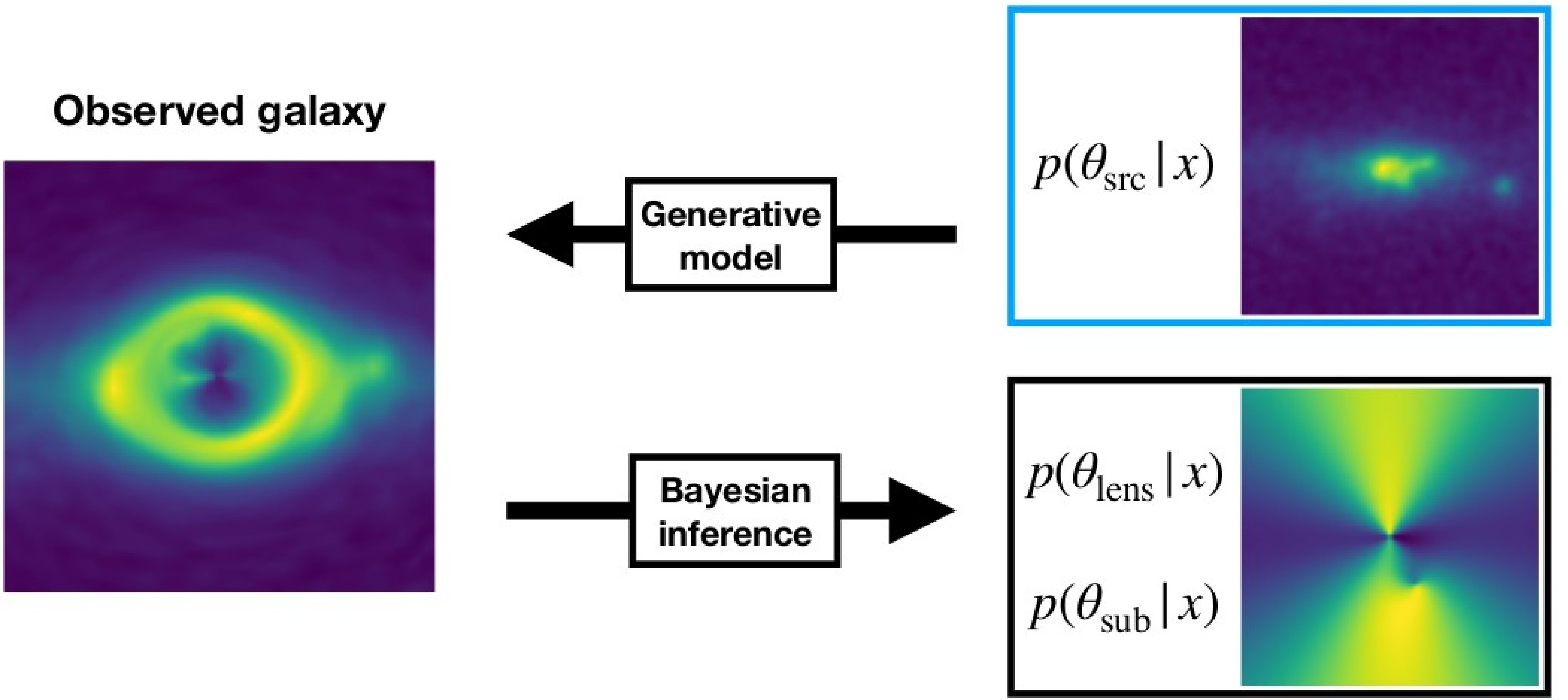
Simple source



We are interested in reconstruction of the displacement field & mass distribution of lens

Slide credit: Adam Coogan

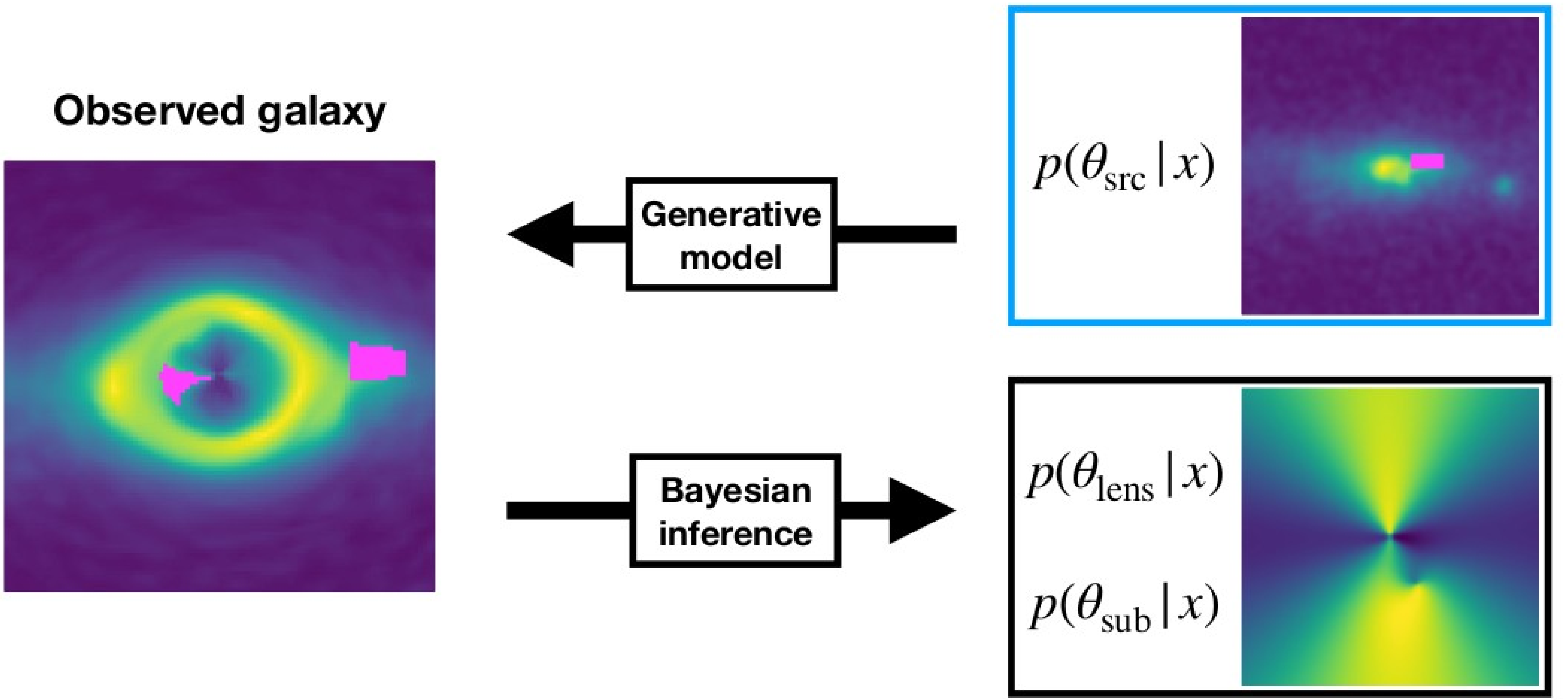
Complex source



Challenge: Sources can be quite complex, with substructure etc.

Slide credit: Adam Coogan

Complex source



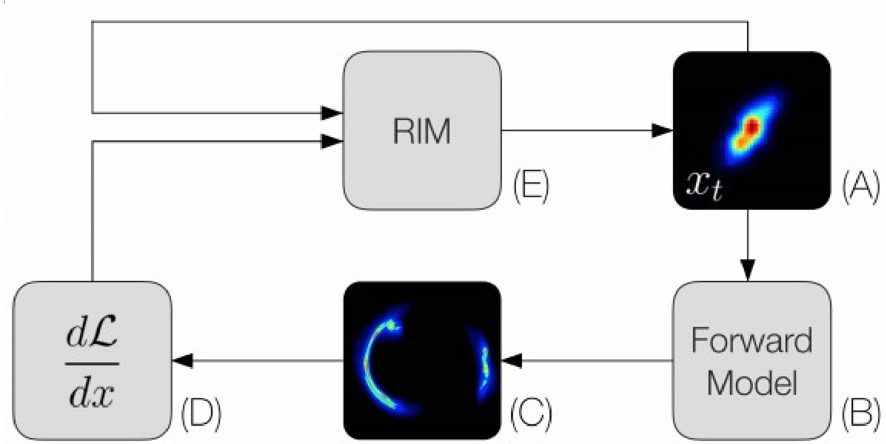
Sources are measured multiple times in image plane (which makes it possible to search for subhalos)

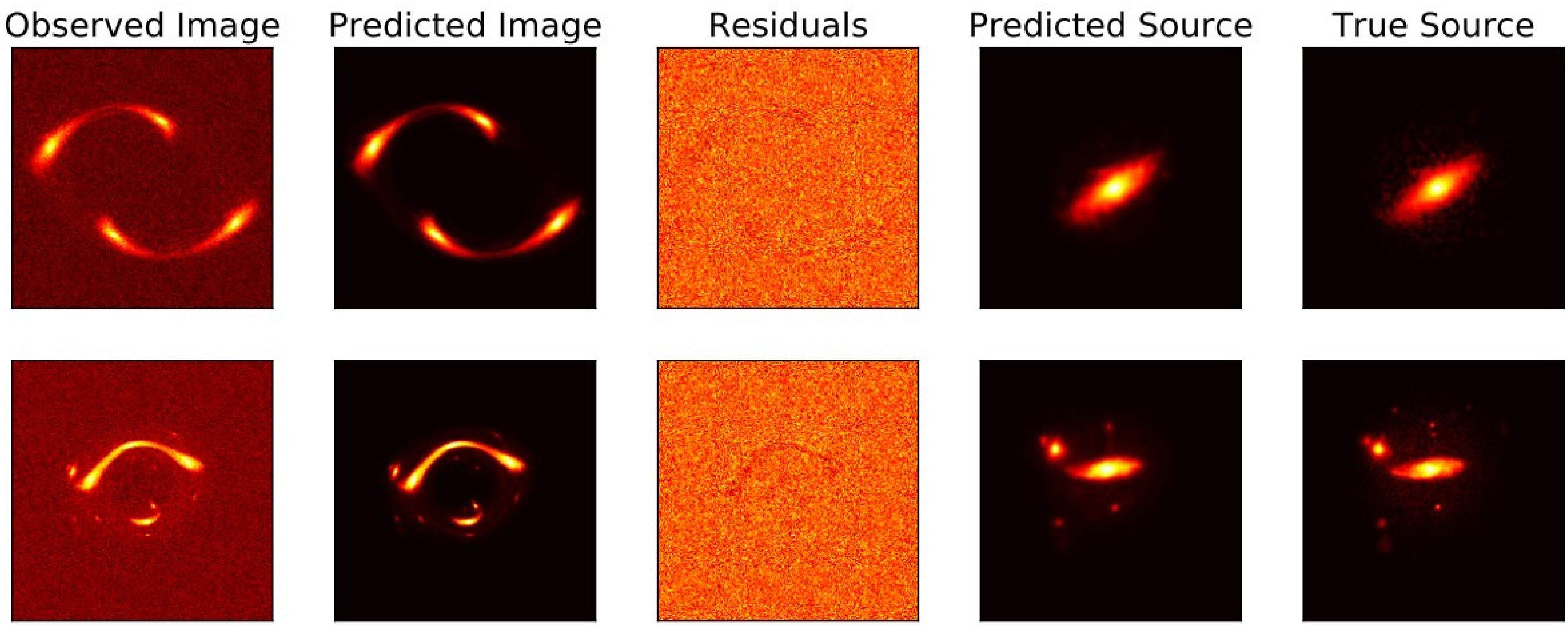
Slide credit: Adam Coogan

Some recent examles with CNNs

Recurrent inference machines for source modeling

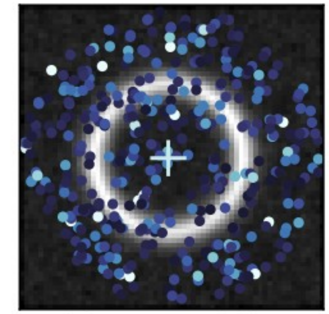
Morningstar+ 2019, https://arxiv.org/abs/1901.01359





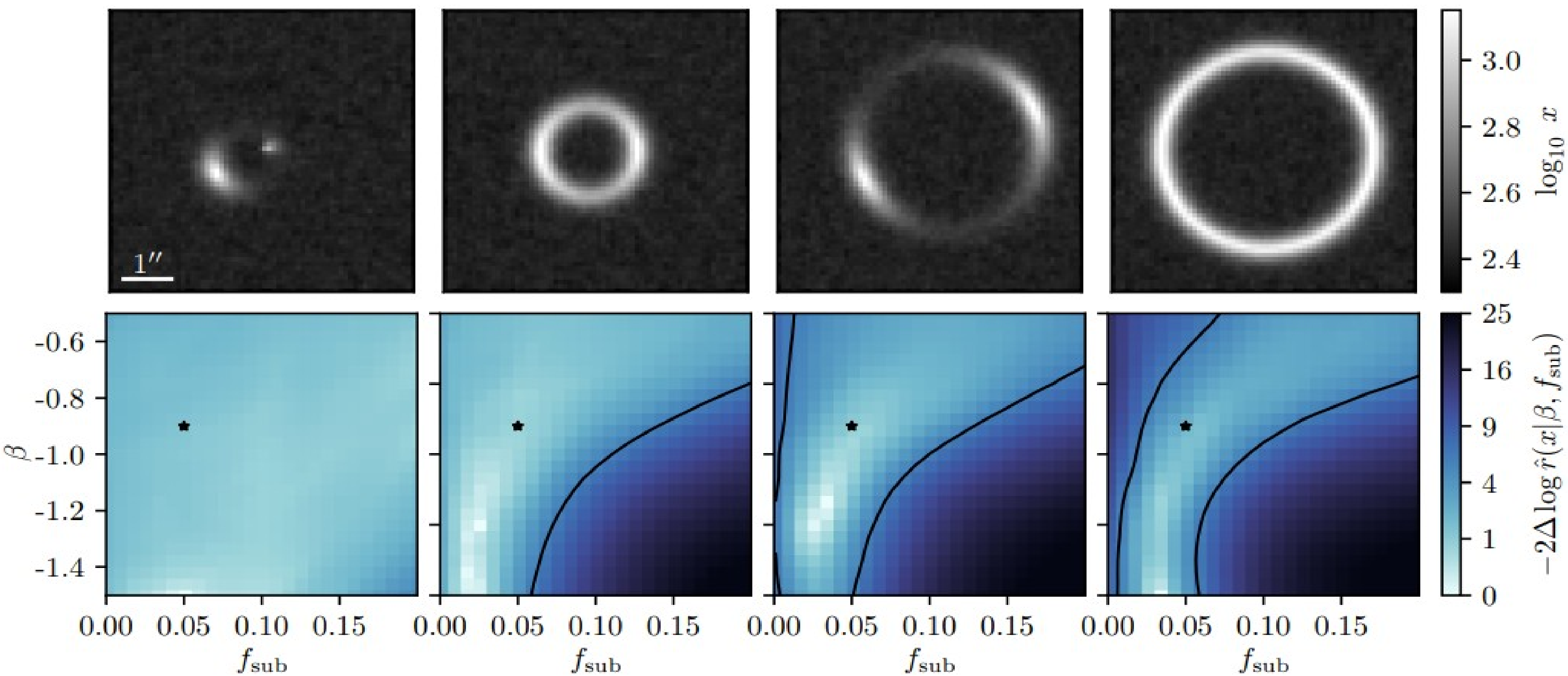
Some recent examles with CNNs

Likelihood free inference for modeling lenses with substructure
Brehmer+ 2019, https://arxiv.org/abs/1909.02005



Use CNN to estimate likelihood ratio*

$$r(x|\theta_0,\theta_1) \equiv \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

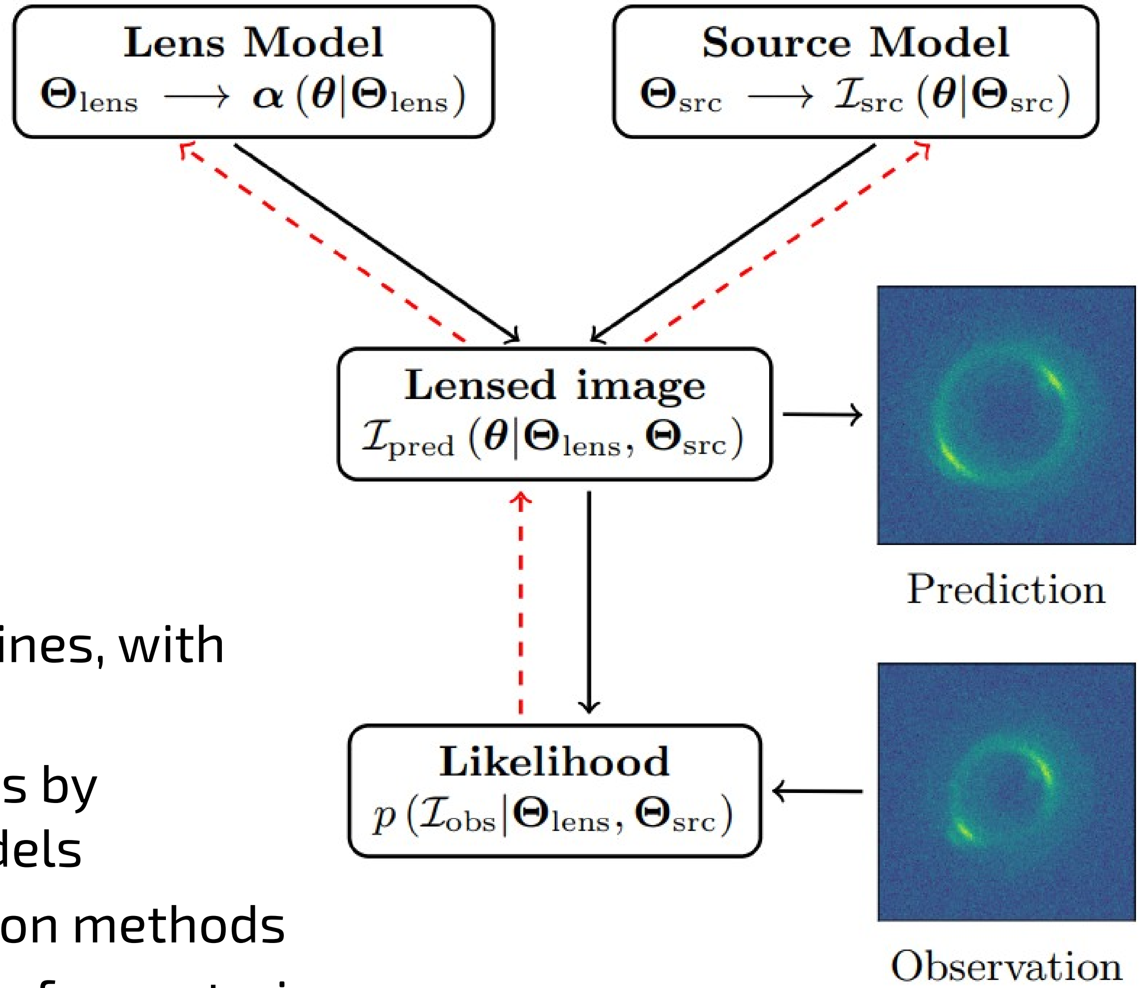


*what is learned is a binary classifier, whose results are then recalibrated to yield a likelihood ratio

Cranmer+ 1506.02169

Our approach

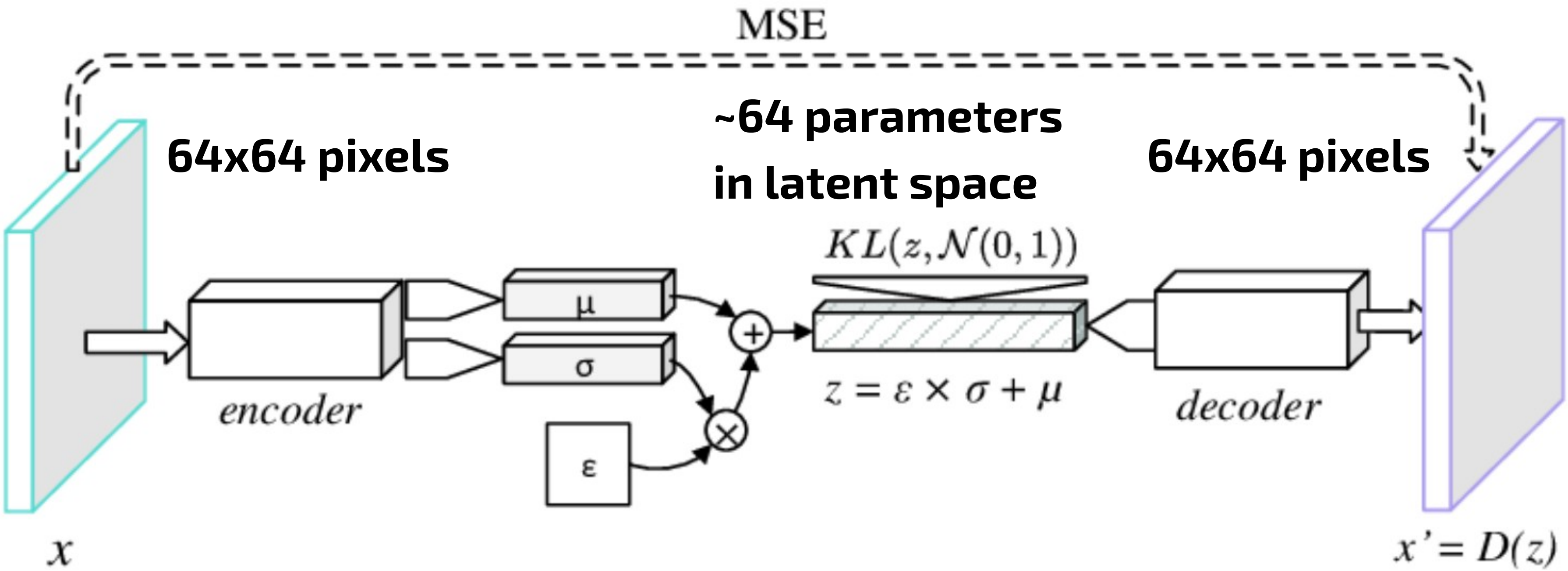
Differentiable probabilistic programming



Philosophy

- Write traditional lensing pipelines, with back-propagation
- Replace individual components by deep/physical generative models
- Use gradient based optimization methods
- Use propabilistic programming for posterior estimates

Variational Auto-Encoder as source model



Components

- Generative model
  $p(x, z) = p(x|z)p(z)$
- Inference model
  $q(z|x)$

Training by ELBO maximization

$$\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{x} \mid \mathbf{z})] - \text{KL}(q(\mathbf{z}) \| p(\mathbf{z}))$$

Marginal log-likelihood

Difference w.r.t. prior

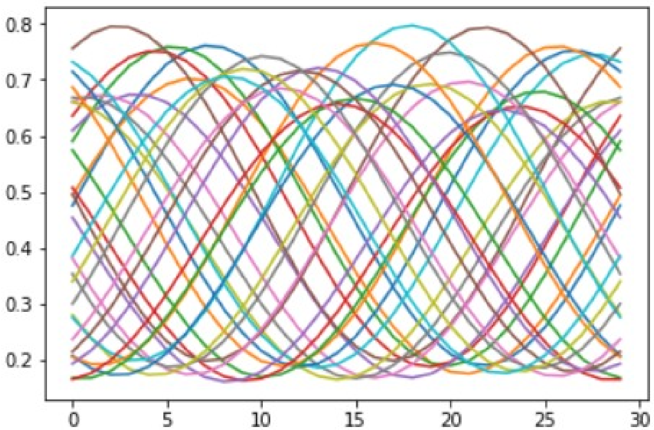
$$\log p(\mathbf{x}) \geq \text{ELBO}(q)$$

Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. arXiv [stat.ML] (2013).

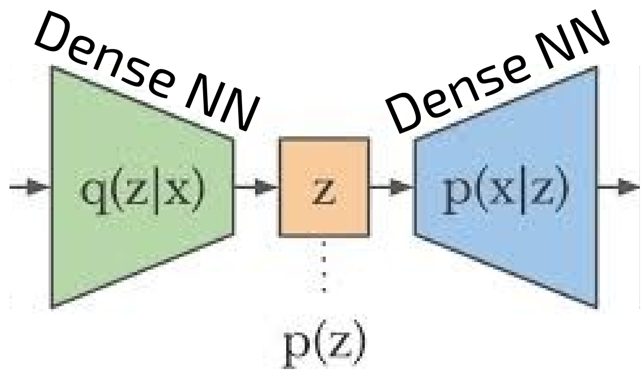
A simple example for the latent space

Training

1000 sine curves

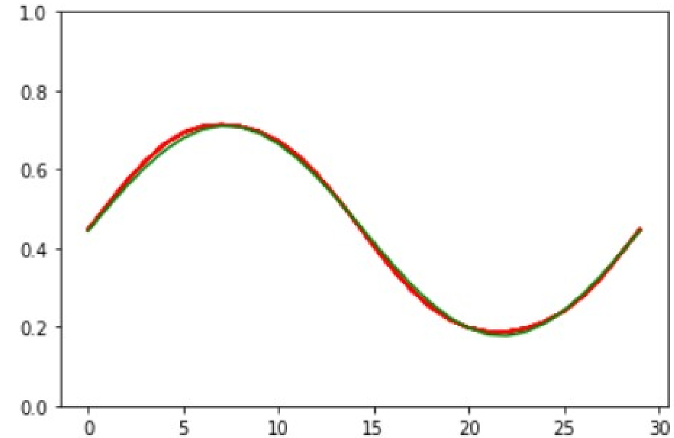


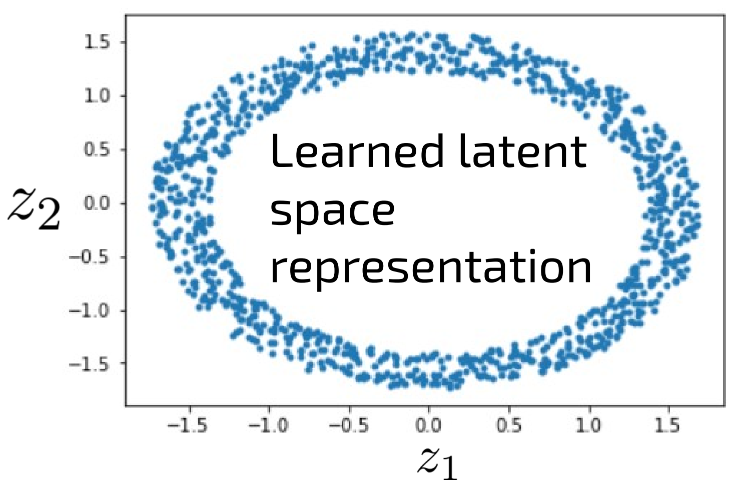
$$f(t) = A \sin(x + \phi)$$



Reconstruction

Works reasonbly well



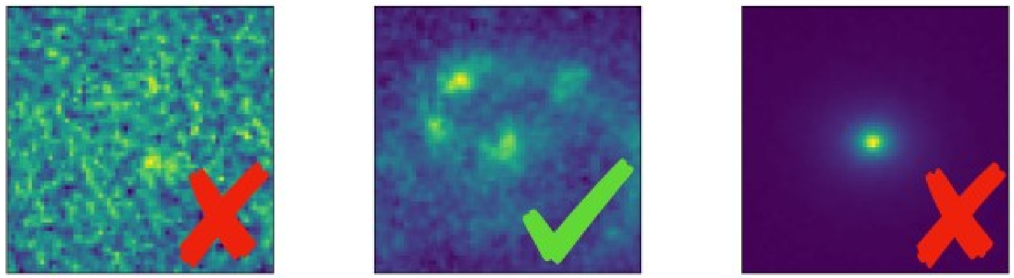


Learned latent space

- Periodic variable (phase)
- Bounded variable (amplitude0

Training data set

- Dataset: ~56,000 galaxies, redshifts ~ 1

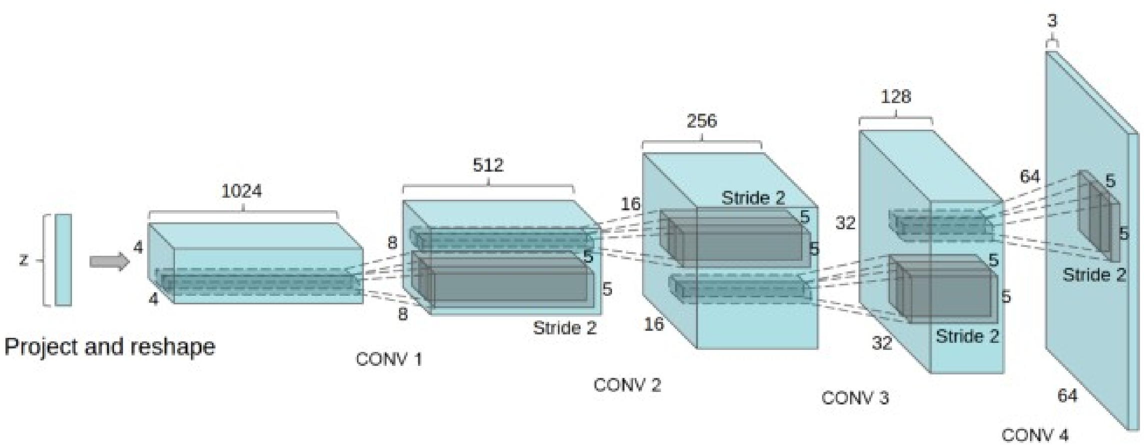


S/N < 10 S/N ~ 20 S/N > 100

This talk: train on ~10,000 images with S/N = 15 - 50

- Encoder, decoder: deep convolutional neural networks

Eg, decoder:



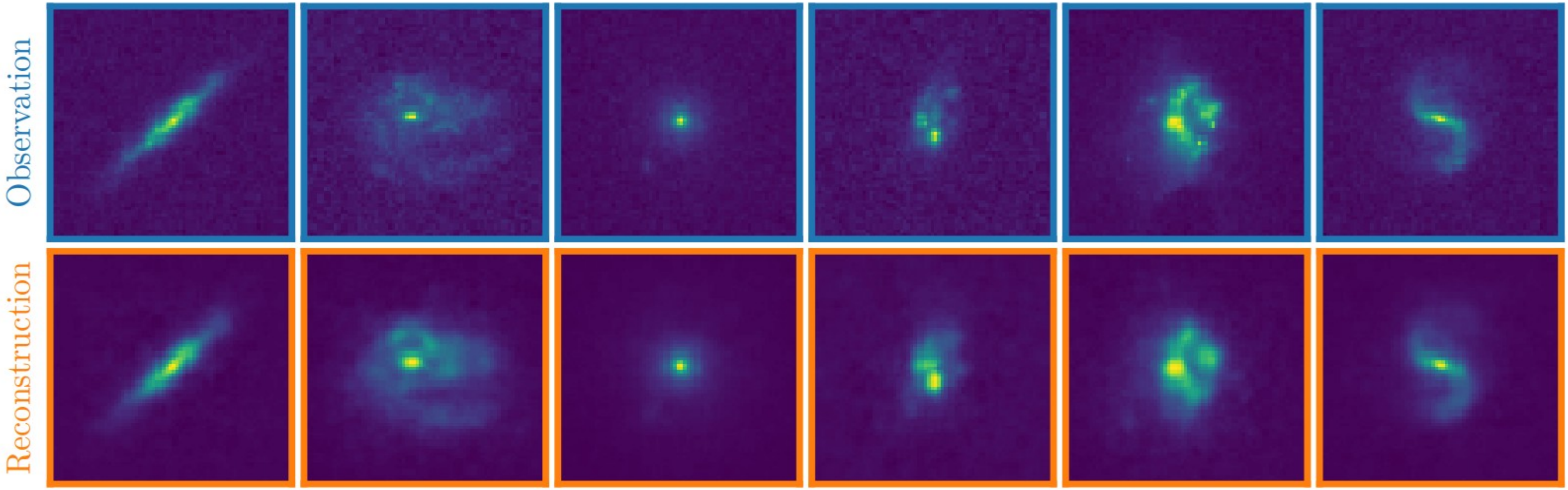
http://great3.jb.man.ac.uk/

Radford et al 2015 (DCGAN)

Slide credit: Adam Coogan

Source galaxy reconstruction w/o lensing

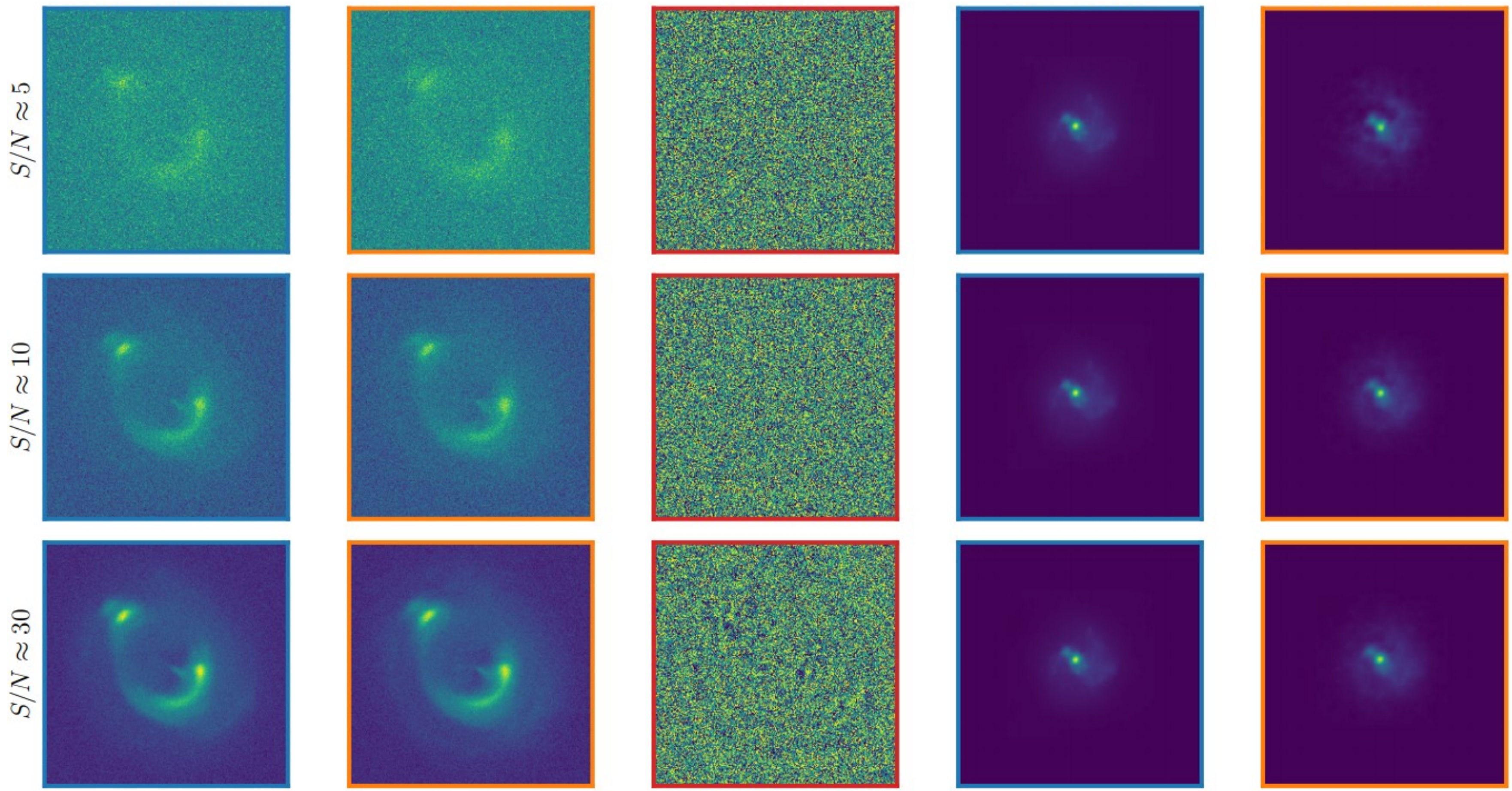
Galaxy image → (Encoder) → Latent space → (Decoder) → Reconstruction



Generative model (or "decoder") seems to be expressive enough to model real galaxies (though somewhat blured).

Can we use this in a "traditional" fit to lensed images?

Source galaxy reconstruction with lensing



Mock data

Residuals

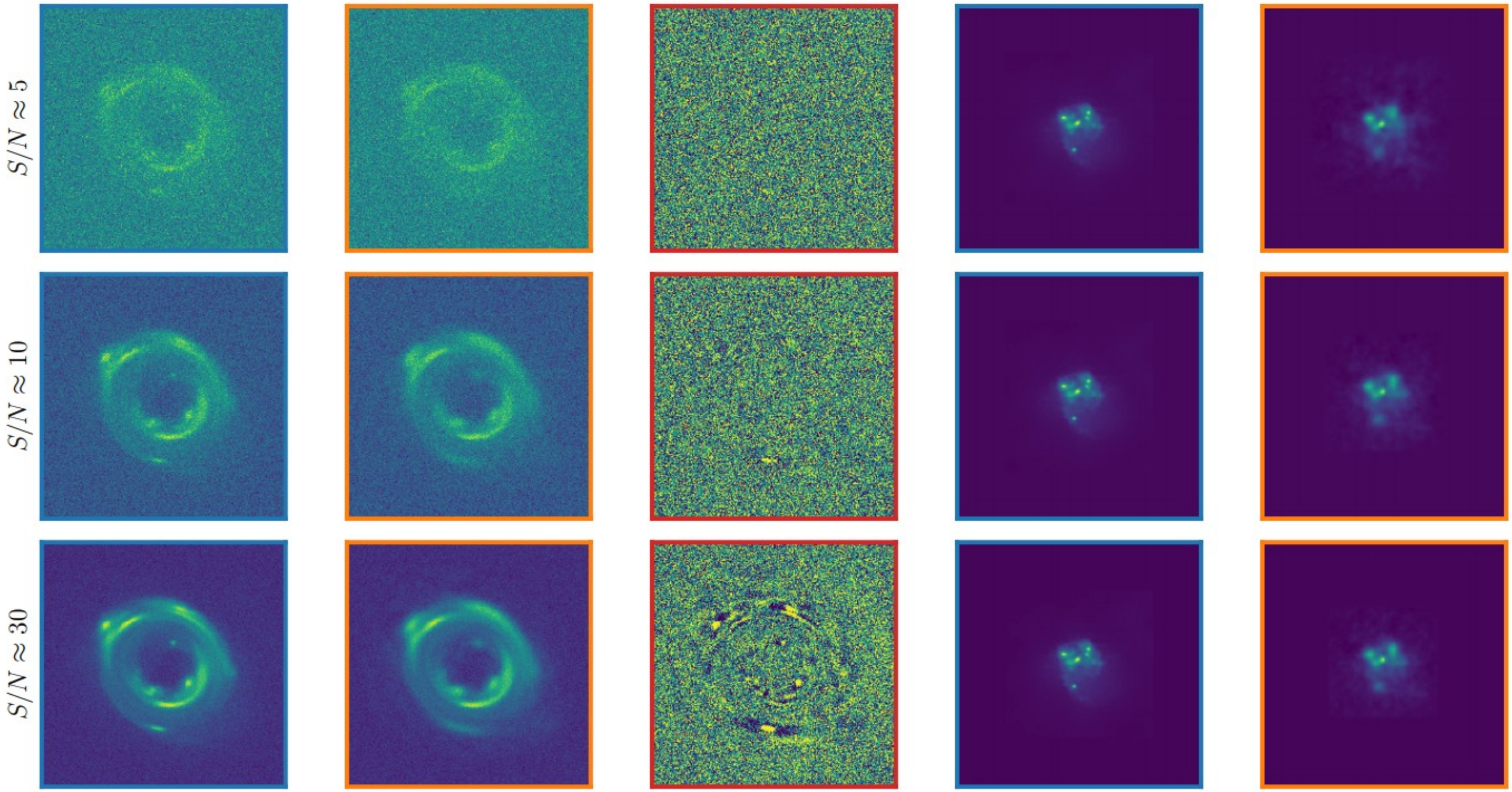
Best-fit source

Original source

Test system 1

Fit with 8 lens parameters
and 64 source paramters

Source galaxy reconstruction with lensing



Mock data

Residuals

Best-fit source

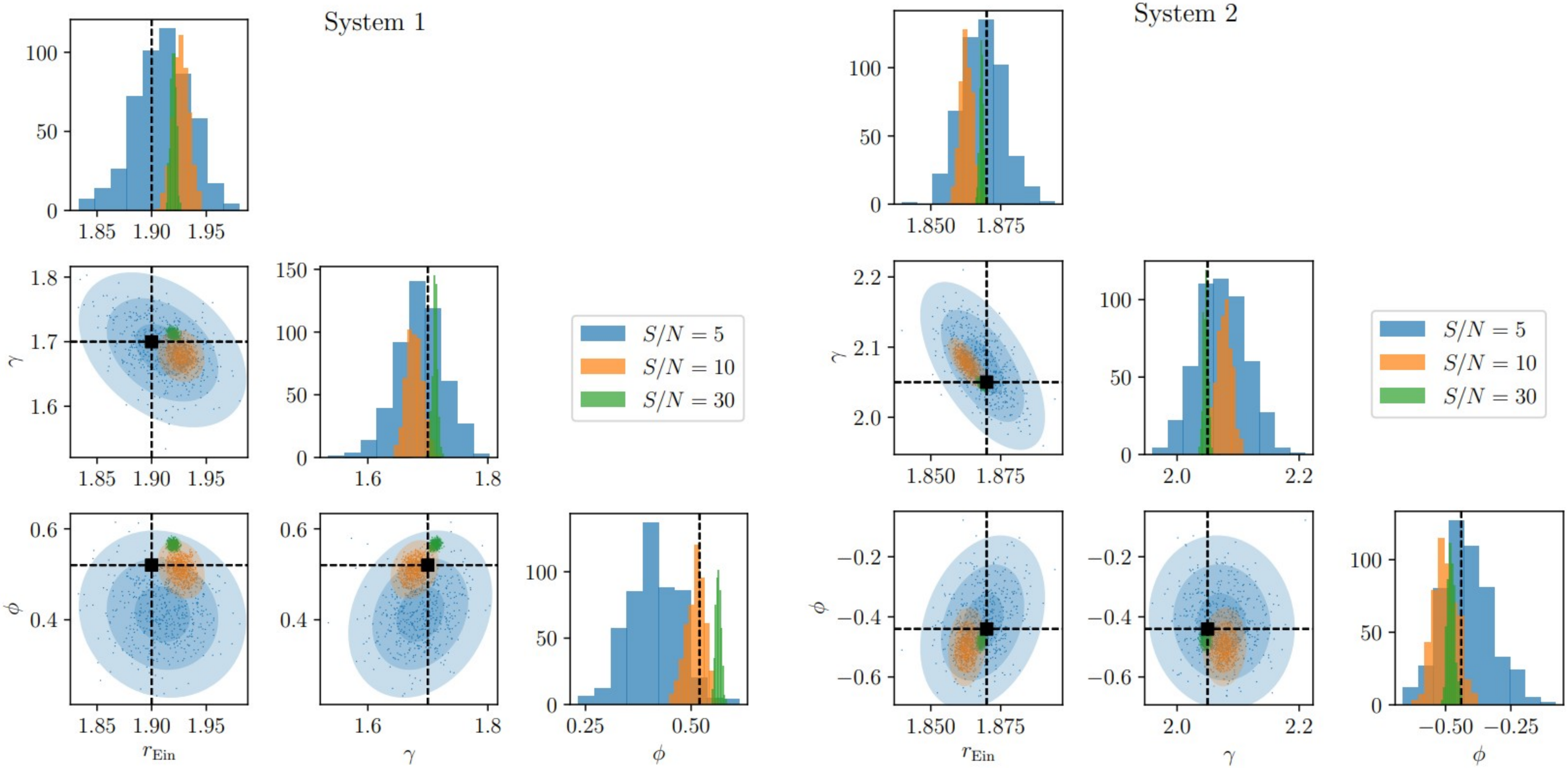
Original source

Test system 2

Fit with 8 lens parameters
and 64 source paramters

Parameter reconstruction with HMC

We use Hamiltonian Monte Carlo to sample the ~75 dimensional posterior.



- Works excellent, but results are slightly biased for how S/N images
- Likely due to limited expressiveness of source model
- Estimating effect on subhalo searches is work in progress

Summary

- Deep neural networks are power flexible function approximators with many applications
- Gradient decent is one of the key components of training neural networks
- Gradients are useful for high-dimensional optimization, sampling and variational inference
- First steps towards gradient-based lensing pipeline that integrates deep generative models look very promising
- Tons of opportunites for improving physics and data analysis, largely unchartered territory

Thank you

Backup slides

The Evidence Lower Bound (ELBO)

We can write:

$$\mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})) = \mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log p(\mathbf{z} \mid \mathbf{x})]$$
$$= \mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] + \log p(\mathbf{x})$$

Model evidence, usually untractable

Define the Evidence Lower Bound:

$$\mathrm{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$$

Since

$$\log p(\mathbf{x}) = \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})) + \mathrm{ELBO}(q)$$

we find

$$\log p(\mathbf{x}) \geq \mathrm{ELBO}(q)$$

Optimizing q*(z) equals maximizing ELBO.

Note that:

$$\mathrm{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z})] + \mathbb{E}[\log p(\mathbf{x} \mid \mathbf{z})] - \mathbb{E}[\log q(\mathbf{z})]$$
$$= \mathbb{E}[\log p(\mathbf{x} \mid \mathbf{z})] - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})).$$

Marginal likelihood　　Difference w.r.t. prior

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians (1601.00670).

Rotation test