Machine Learning in Exoplanet Atmospheric Characterisation

Ariel Conference 2020

Dr Ingo Waldmann

Azure
UK SPACE AGENCY
Horizon 2020
erc European Research Council
Established by the European Commission
Science & Technology Facilities Council

TauREx 3

- Built from the ground up as full python stack
- 10 - 200 times faster than TauREx 2
- Full NVIDIA and OpenCL GPU support (another 50x faster for JWST or high-res)
- Fully tested against TauREx 2 which is benchmarked against NEMESIS, CHIMERA, ARCiS
- For full installation type: “pip install taurex”
- Plugin features and TauREx extensions
- New and fast cross sections
- Fully open under BSD license

Molecules	τ-REx 2 xsec (s)	τ-REx 2 k-tables (s)	τ-REx 3 xsec (s)
1	7.23	0.45	0.61
2	8.90	0.78	0.74
4	12.42	1.49	0.92
7	19.02	2.63	1.23
15	263.56	8.21	2.34

Al-Rafaie et al. submitted, arXiv: 1912:07759

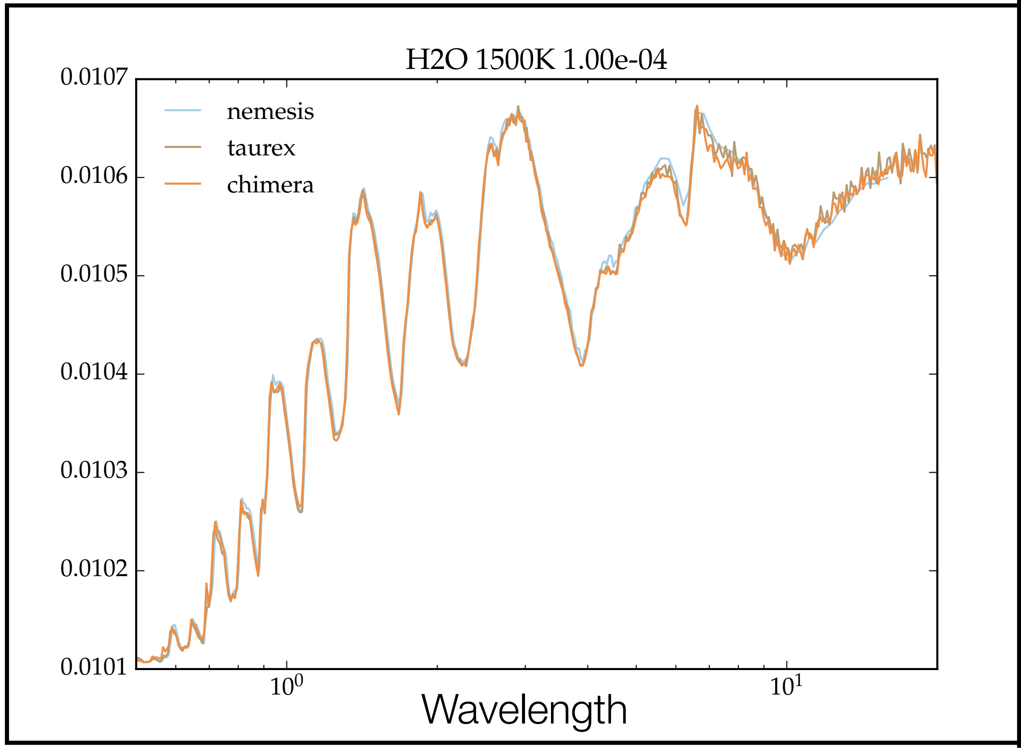
https://github.com/ucl-exoplanets/TauREx3

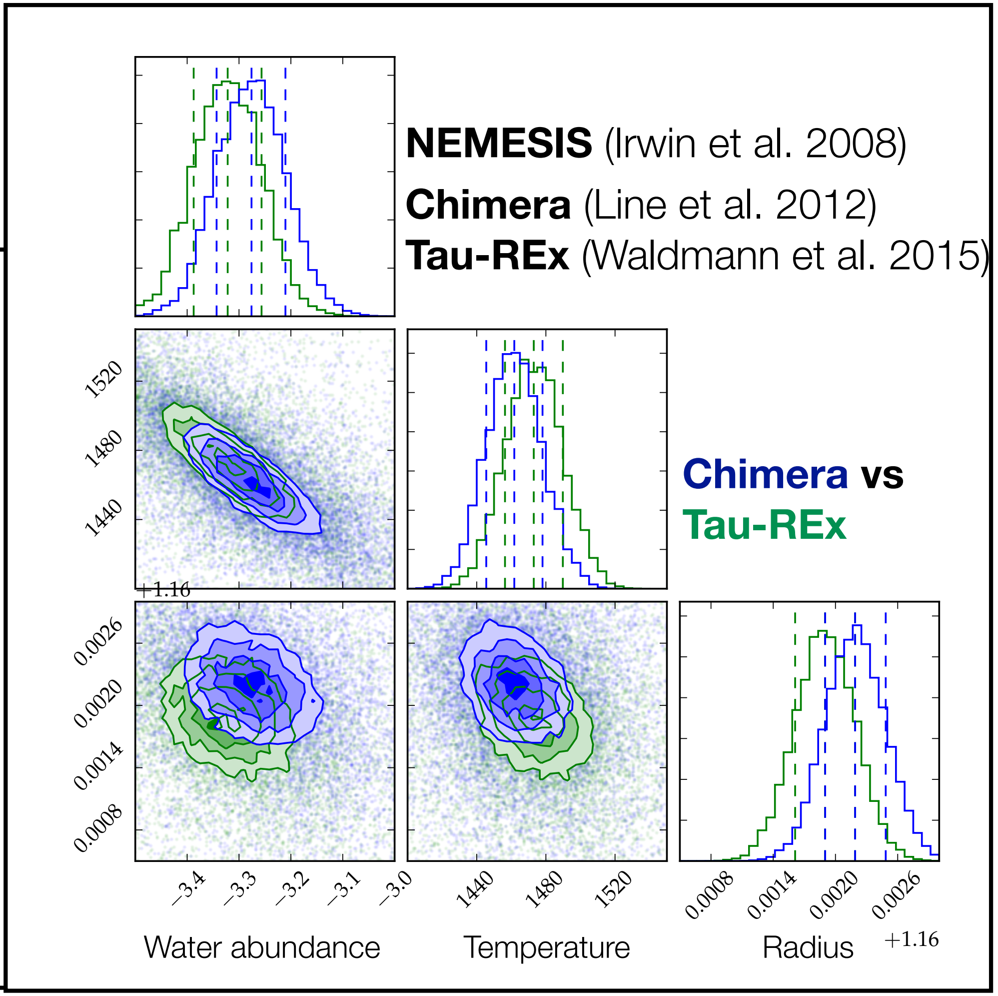
Retrieval model comparison

- We are comparing forward models and retrieval results with Mike Line and Jo Barstow
- Exact comparison between line list differences
- K-coeffcients (NEMESIS) vs cross-section approaches (Chimera, Tau-REx)

- Open up wider model comparison in data challenge later on this year

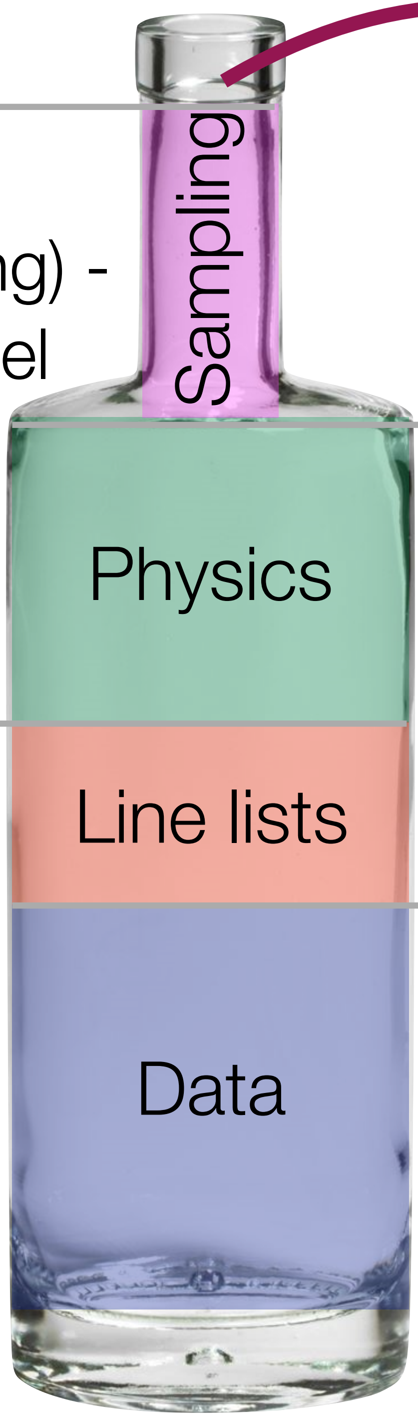
Chimera vs NEMESIS vs TauREx





The retrieval bottleneck

- Classical sampling slow (MCMC, Nested Sampling) -> 10^5 - 10^6 forward model iterations



well balanced paper

- Temperature-Pressure profiles
- Cloud models
- Disequilibrium chemistry
- 3D effects -> GCMs

- Extremely large databases

ExoMol

HiTemp

Hitran

JWST

E-ELT

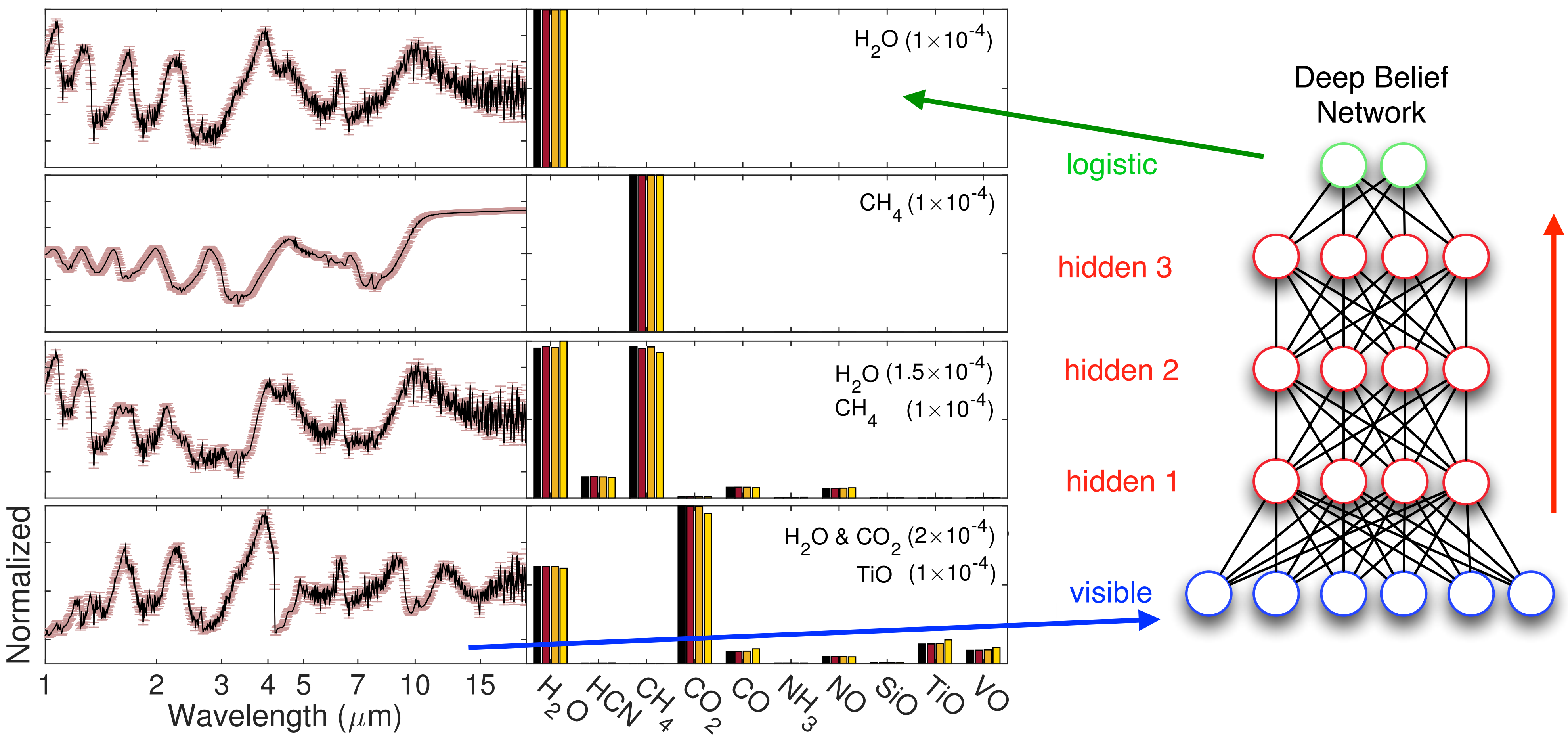
Ariel

Twinkle

Ground Based

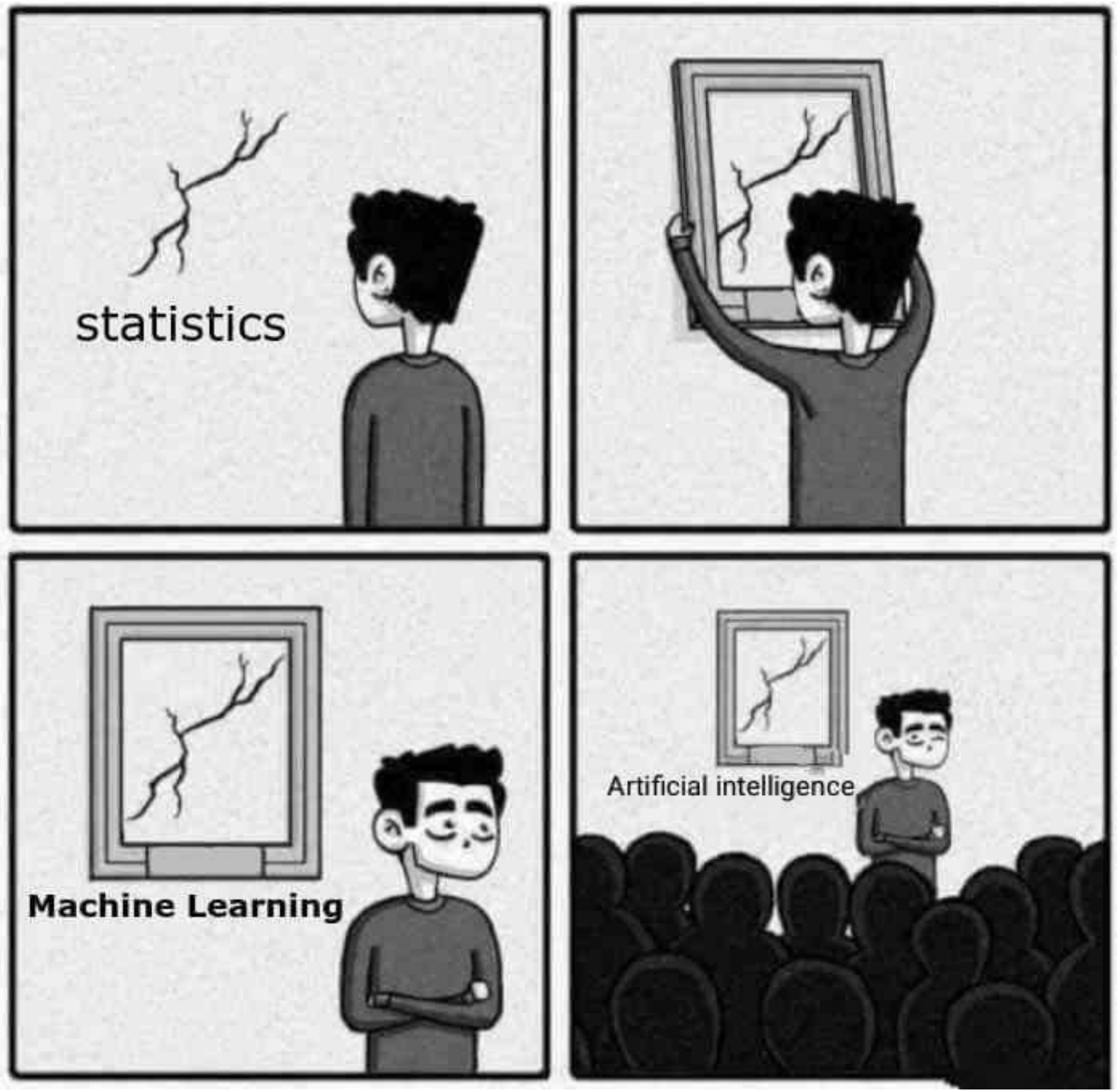
CHEOPS

Robotic Exoplanet Recognition (RobERt)

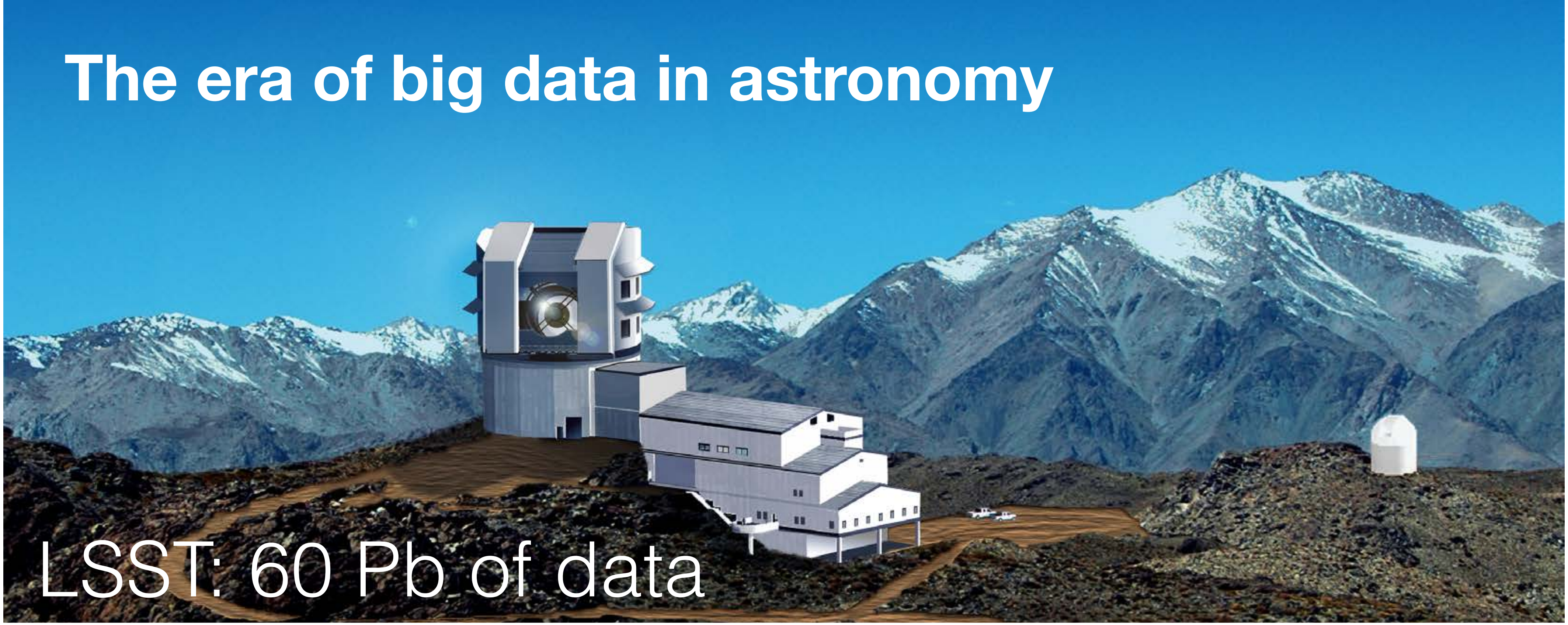


Waldmann 2016

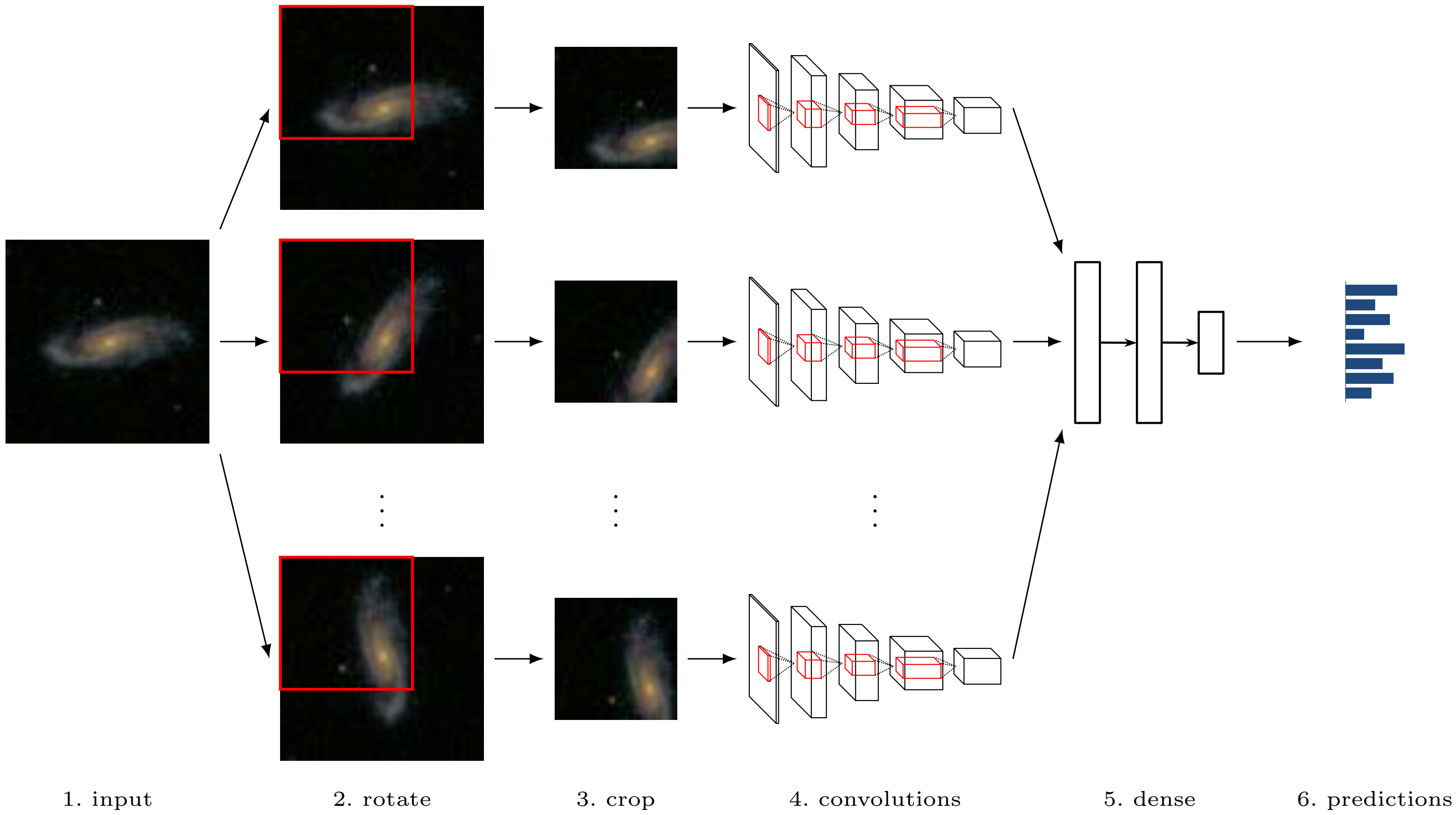
The brave new world of deep learning



The era of big data in astronomy

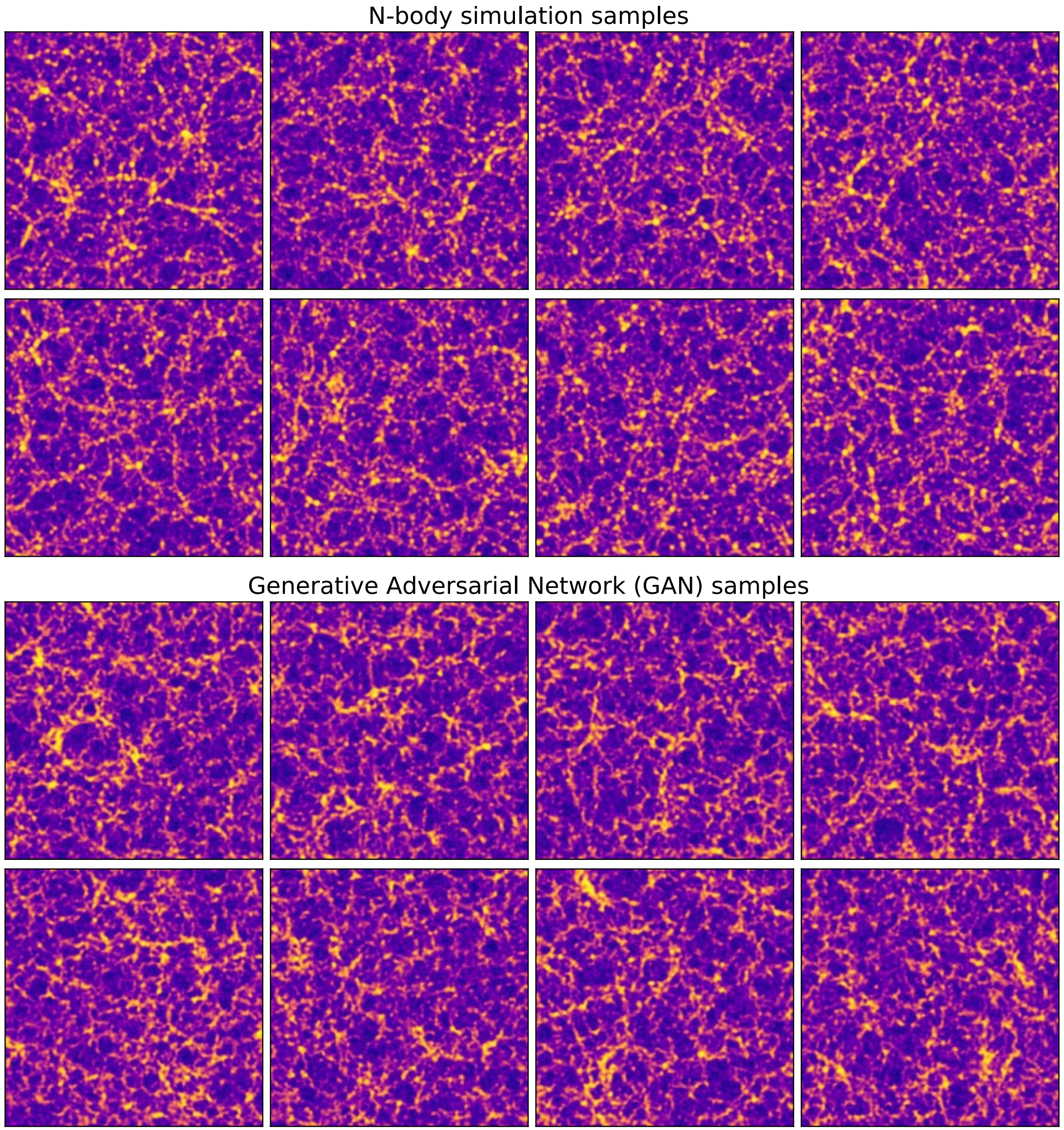


Classifying galaxies in Galaxy Zoo



e.g. Dielmann et al. 2015, Lukic et al. 2018

Learning the cosmic web from N-body simulations



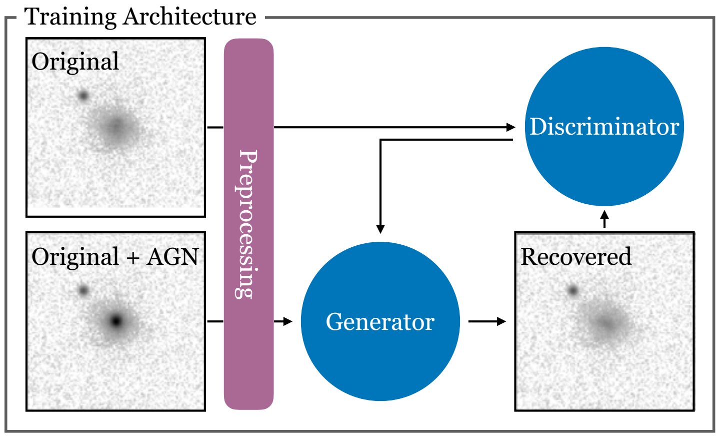
Rodriguez et al. 2018

Some other application examples

- Superresolution imaging of planetary surfaces
- De-trending in weak lensing
- Crater counting on planetary surfaces
- Learning instrument responses
- etc

PSFGAN

Learning instrument point spread functions from data



Stark et al. 2018

Searching for exoplanets

- The Kepler and TESS data set is ideal to train neural networks
- Neural nets can outperform more classical detection pipelines
- Can probe lower signal-to-noise data than other methods
- Can include domain knowledge in search (Ansdell et al. 2019)

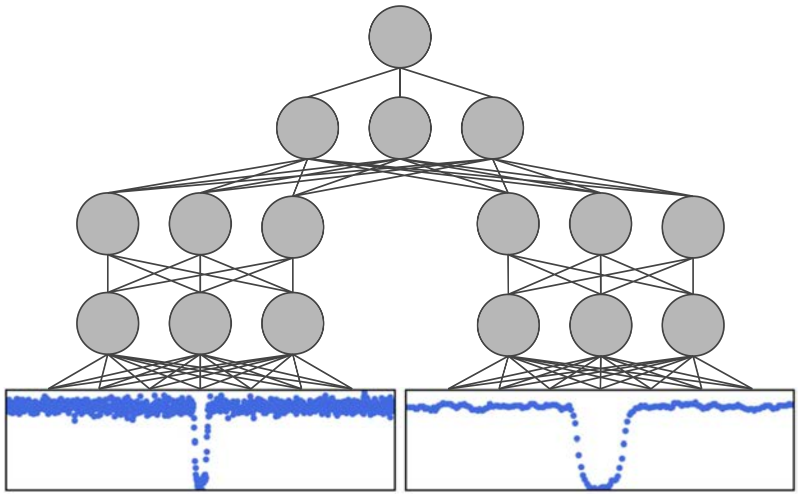
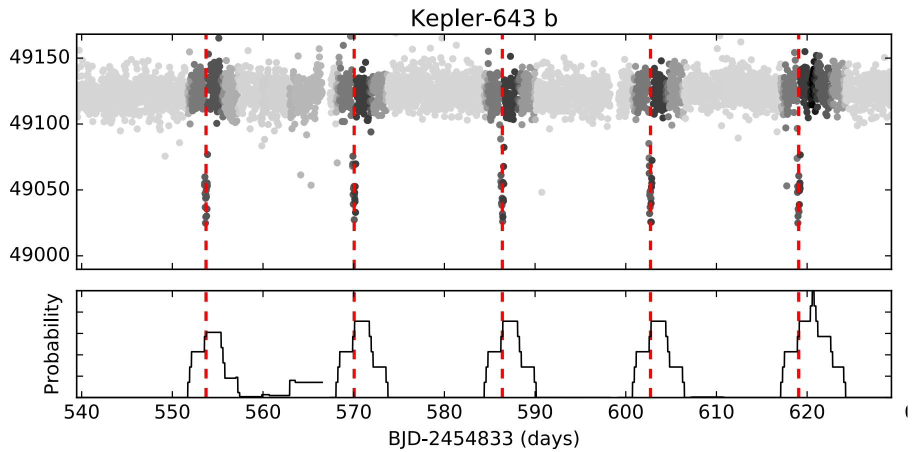


Figure 4. Fully connected neural network architecture for classifying light curves, with both global and local input views.



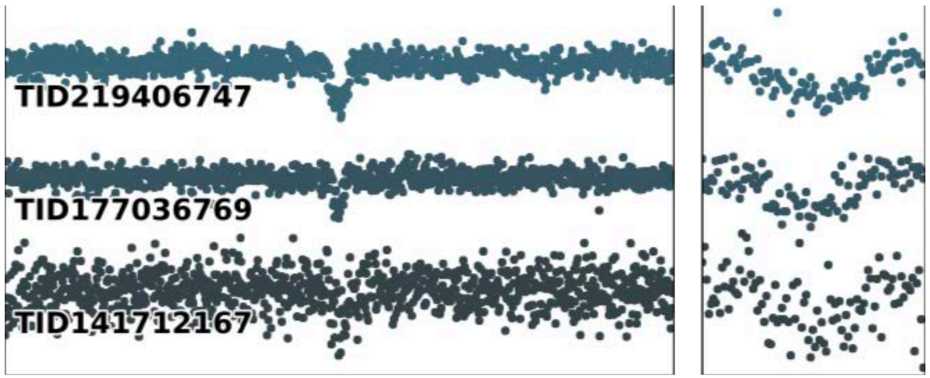
Shallue & Vandenburg 2018, Dattilo et al. 2019

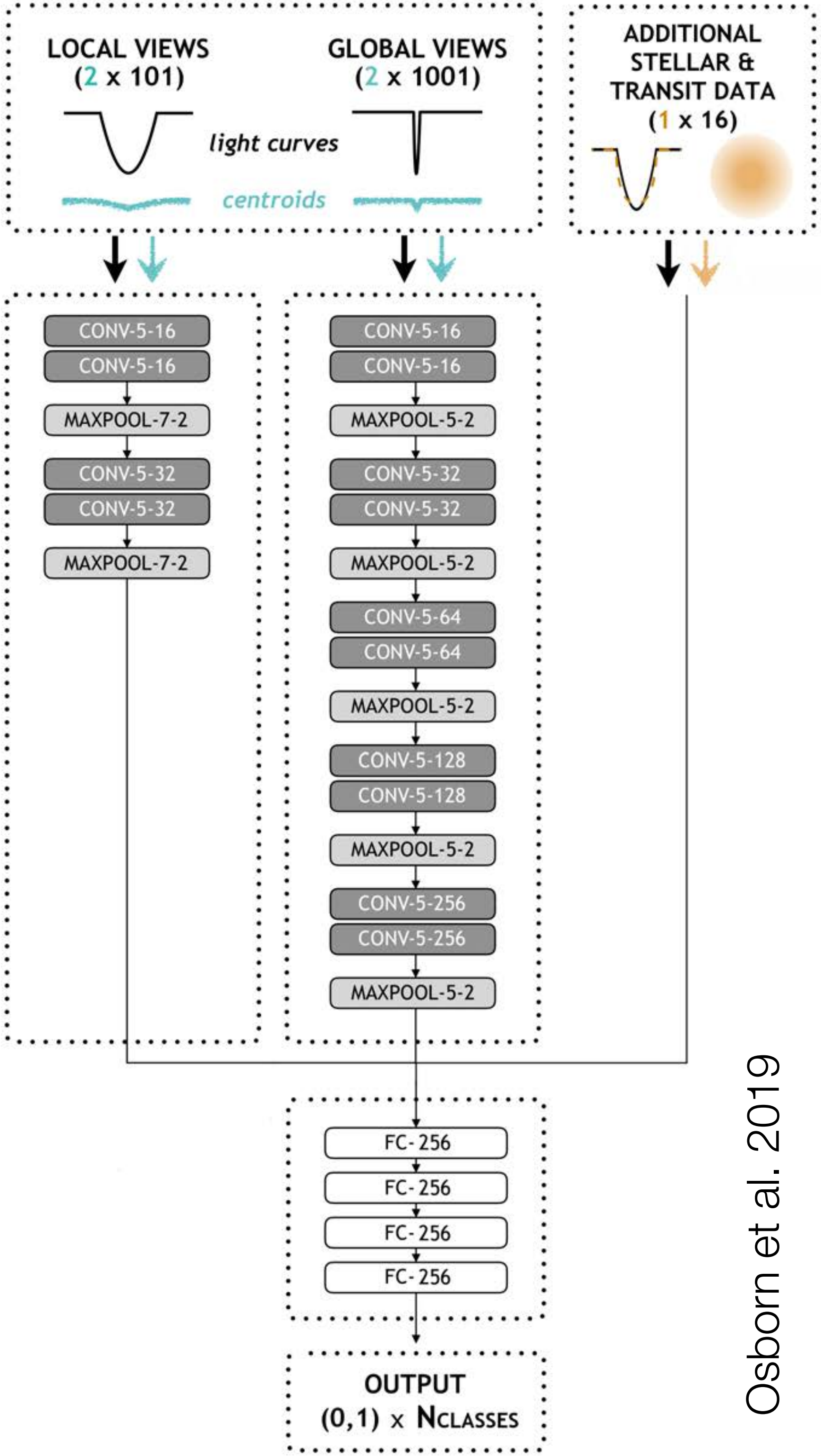
Pearson, Palofax & Griffith 2018

Searching for exoplanets

- Significant work being done in this field using a range of techniques
- Understanding instrument systematics are the main hindrance (pixel sensitivity variations as function of space craft orbit)
- All data is publicly available and well documented

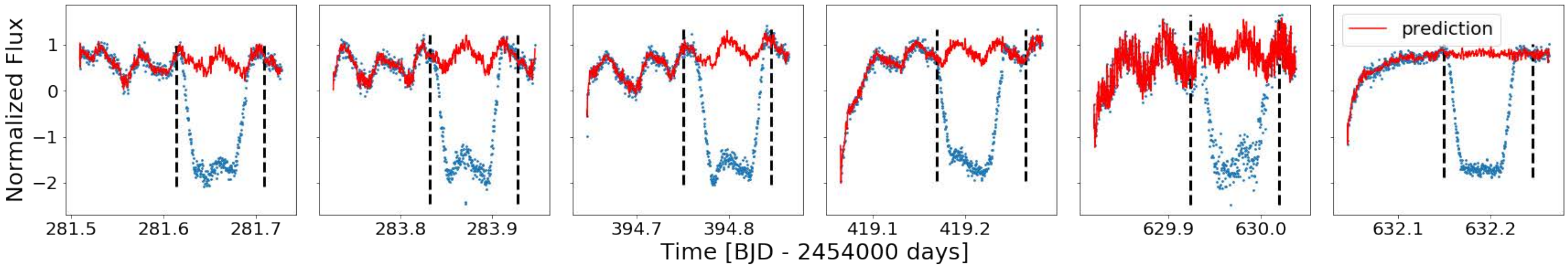
Example of transit data





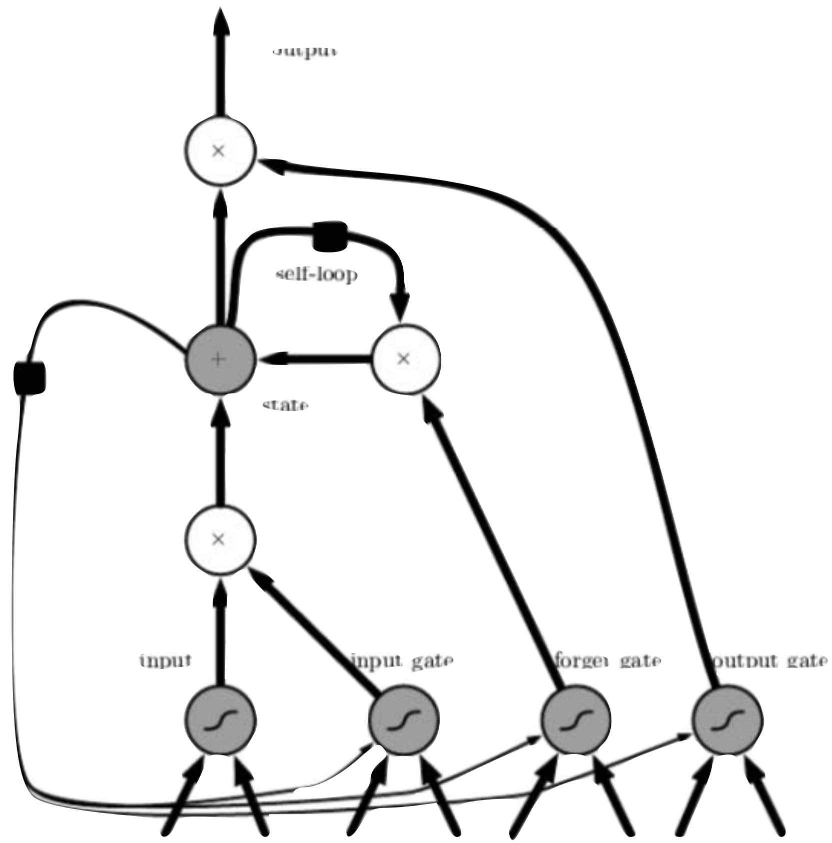
Osborn et al. 2019

Correcting time series data using probabilistic LSTMs



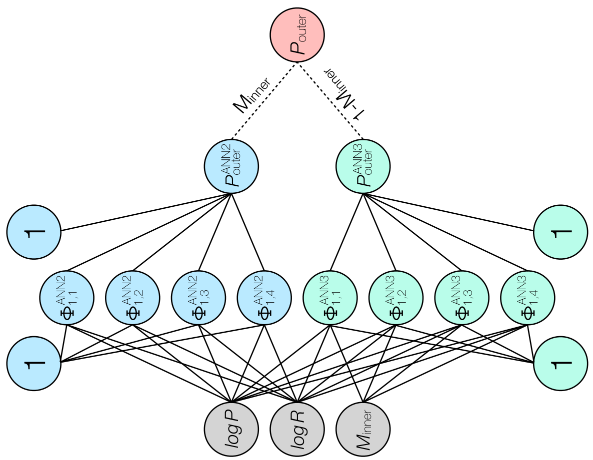
- Exploits the spatial and time dependence of systematic noise
- Nearly infinitely scalable to huge data sets such as Kepler or TESS (Gaussian Processes are restricted here)
- Probabilistic time forecasting, i.e. accurate error bars

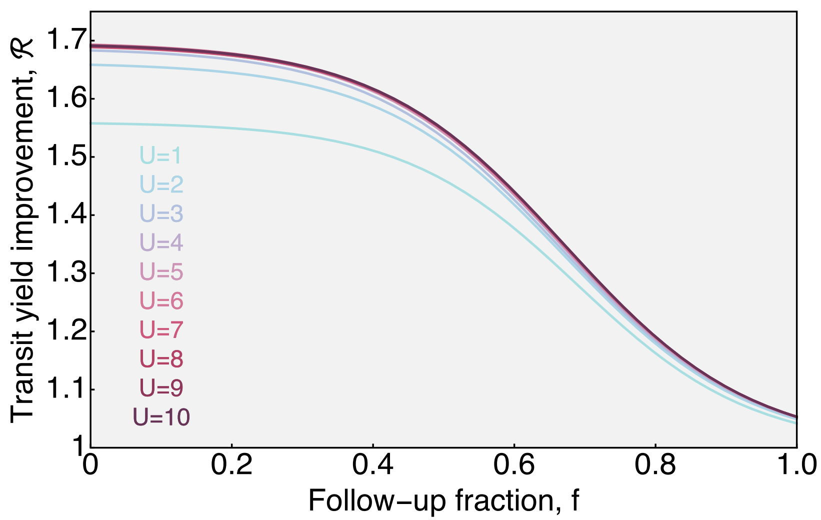
Morvan et al., AJ accepted,
arXiv: 2001:03370



Predicting additional planets in a system

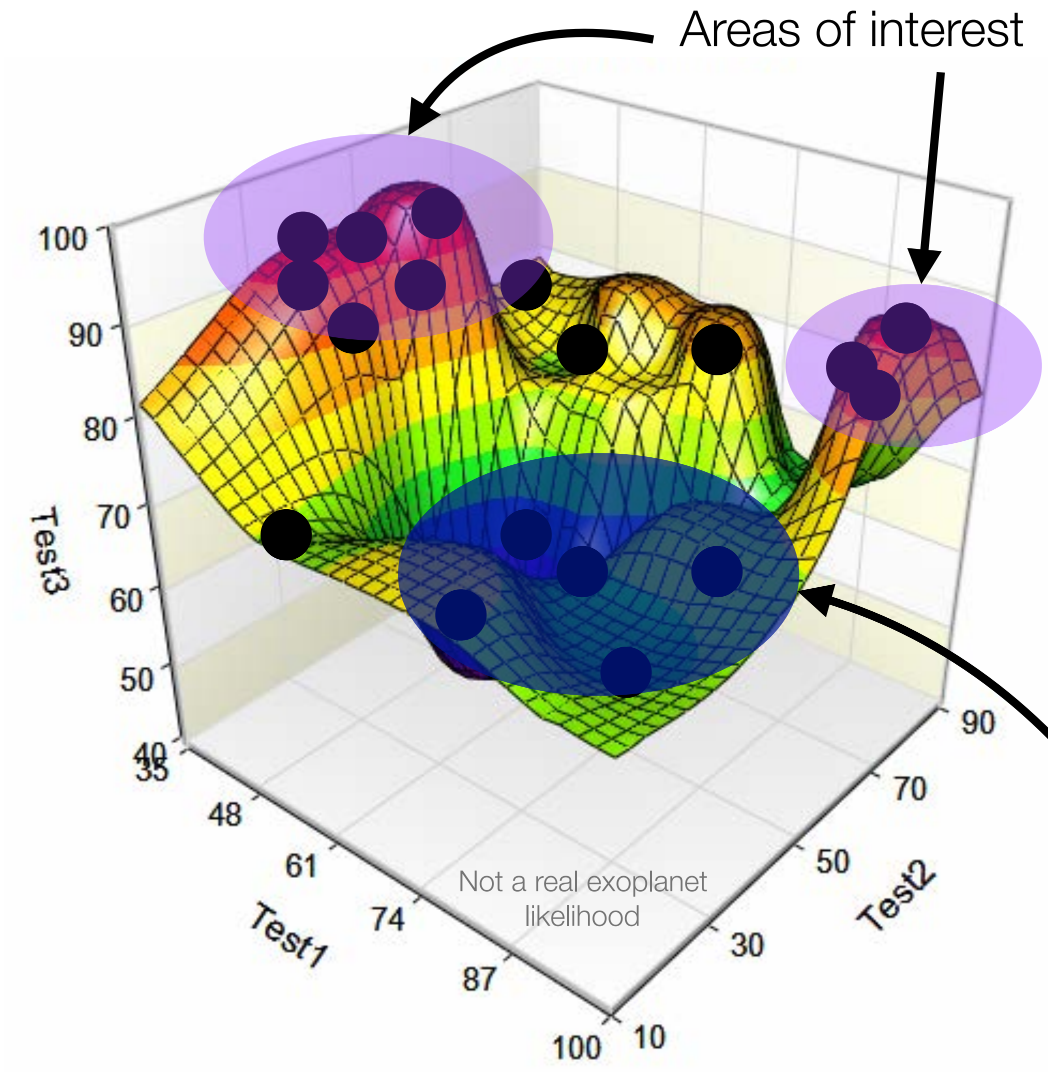
- Predicting the likelihood of extra planets existing in a system
- Potentially increases transit detection efficiencies significantly





Kipping & Lam 2017

Knowing the retrieval likelihood



Sampling is hard and time consuming

Nested Sampling is currently the norm, Drawing “nests” of interest where sampling will be more pronounced.

NS does sample the full likelihood and gets the Evidence.

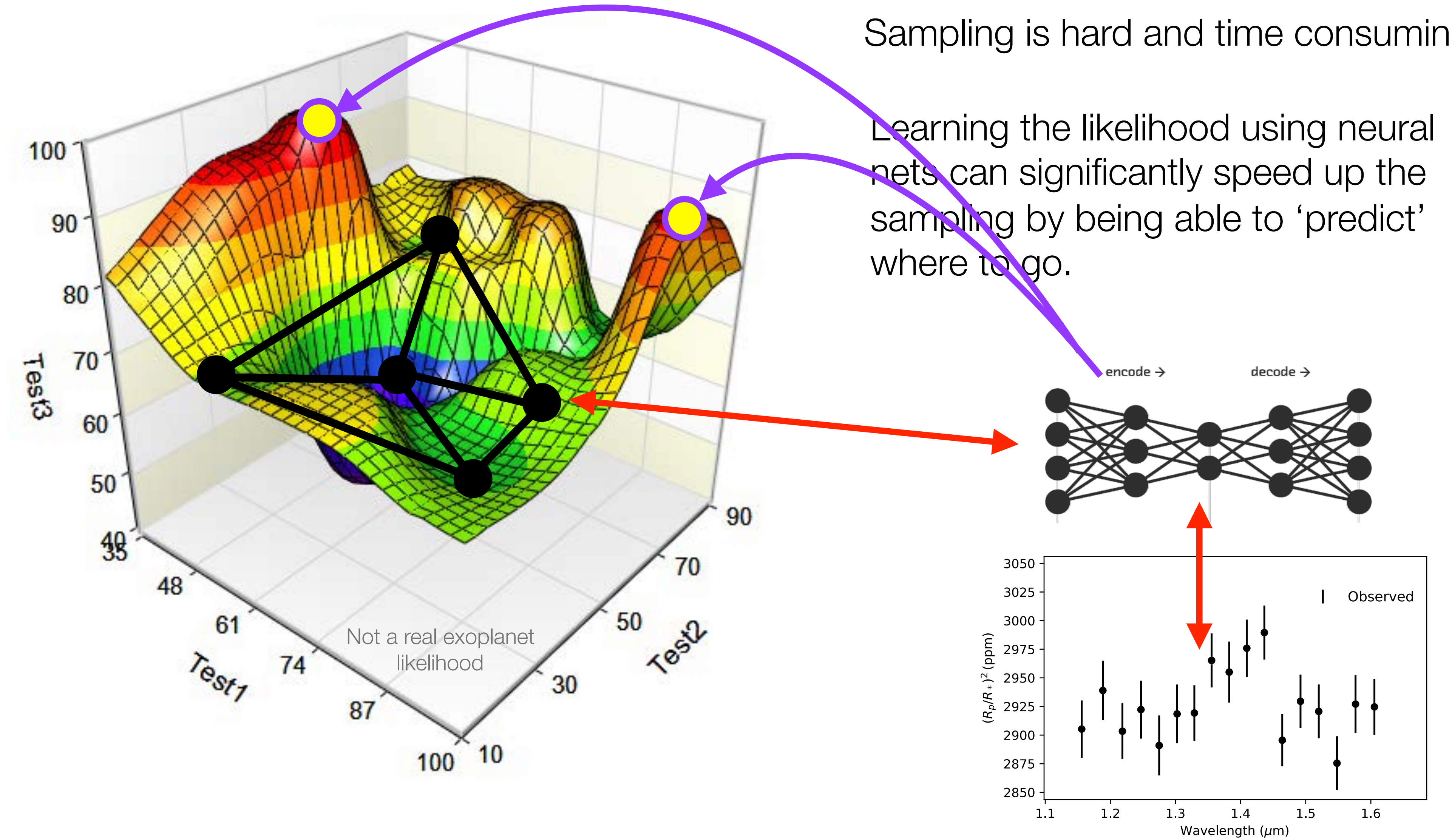
$$p(x) = \int p(x \mid \theta) p(\theta) d\theta$$

Do we need to sample here?

Knowing the retrieval likelihood

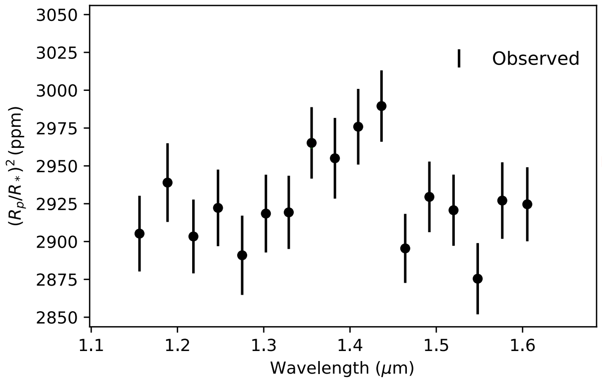
Sampling is hard and time consuming

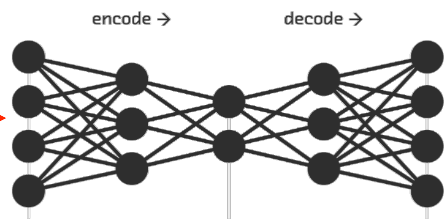
Learning the likelihood using neural nets can significantly speed up the sampling by being able to ‘predict’ where to go.

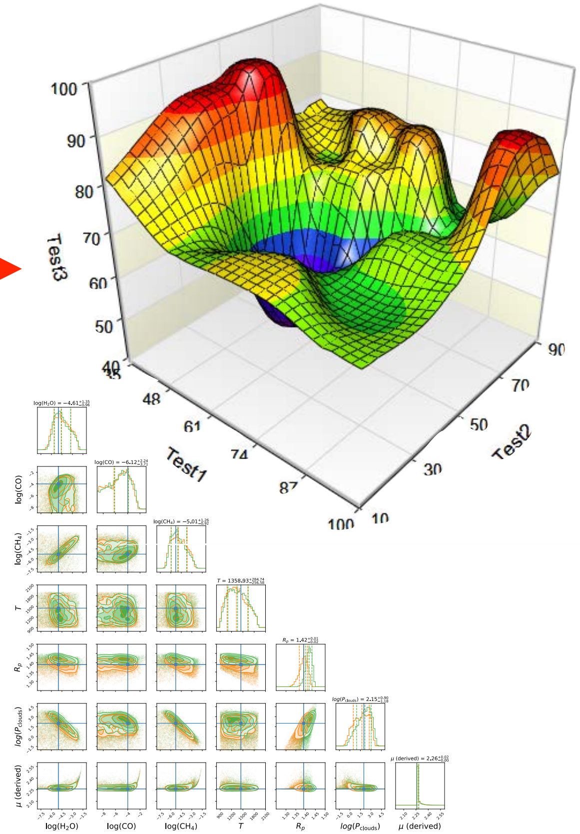


Producing the mapping between data and posterior

Can we map from the data to the posteriors directly?

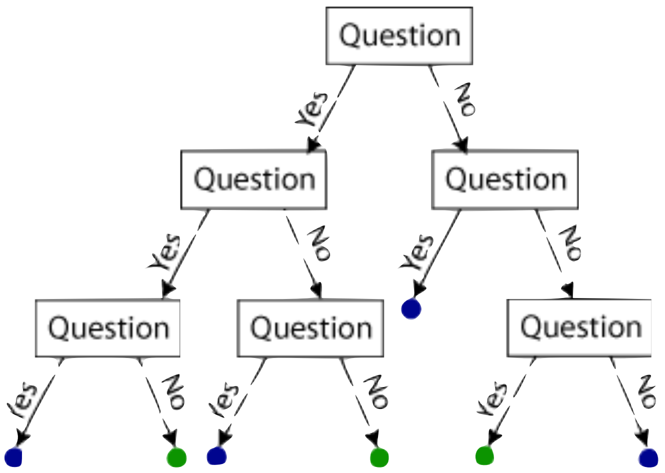


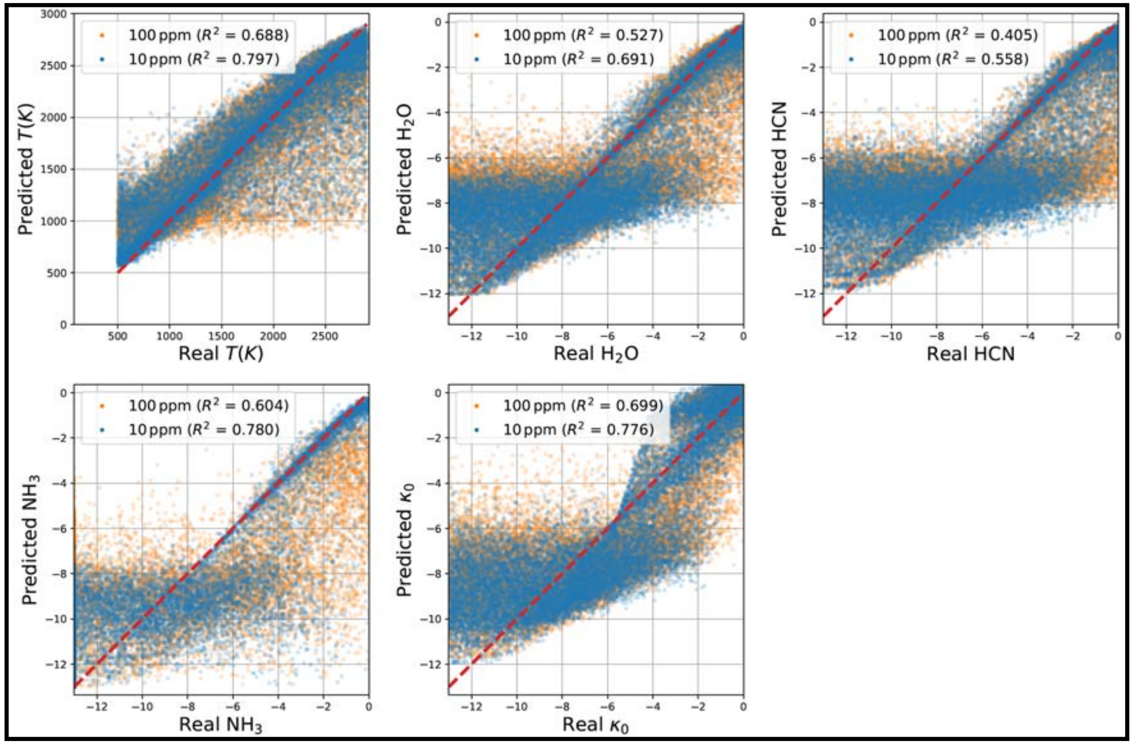




Machine learning atmospheric retrievals

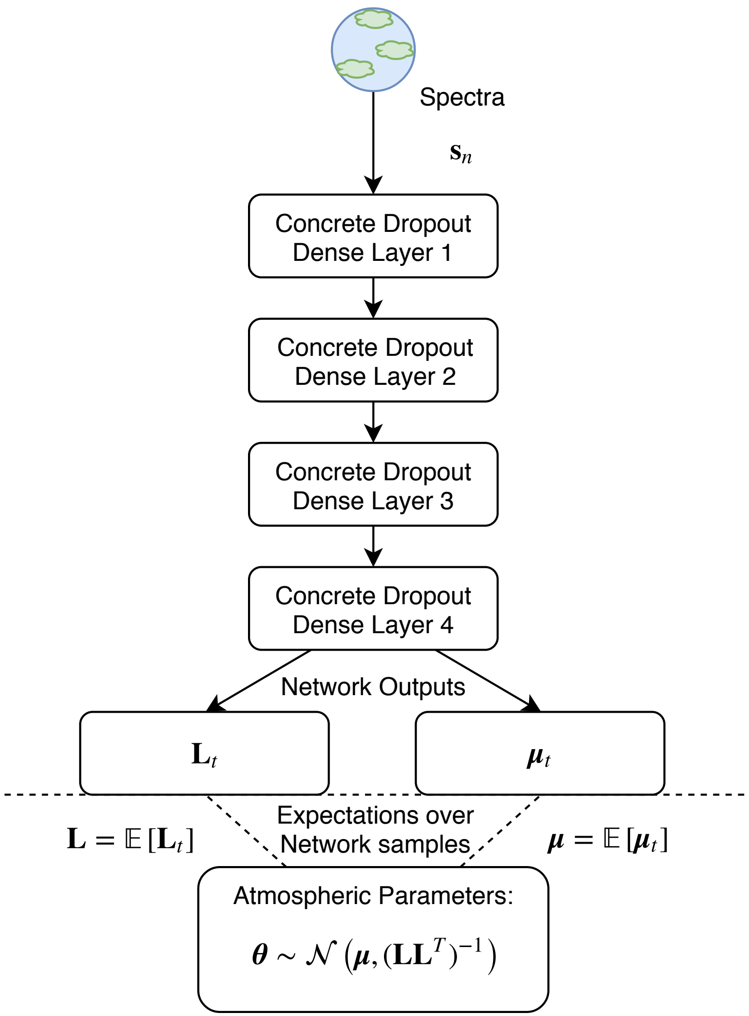
- Machine learning approach using random forests
- Learns to repeat retrieval of a planet (e.g. WASP-12b) very fast





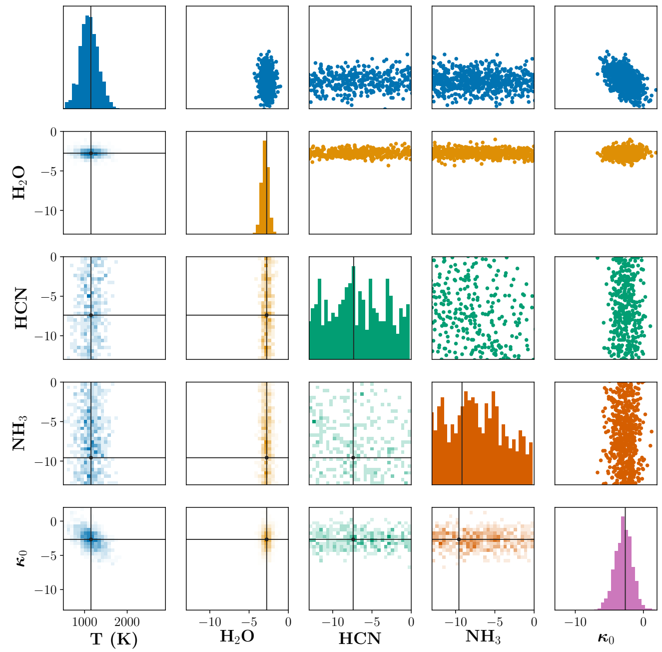
Márquez-Neila, Fisher, et al. 2018

Machine learning atmospheric retrievals



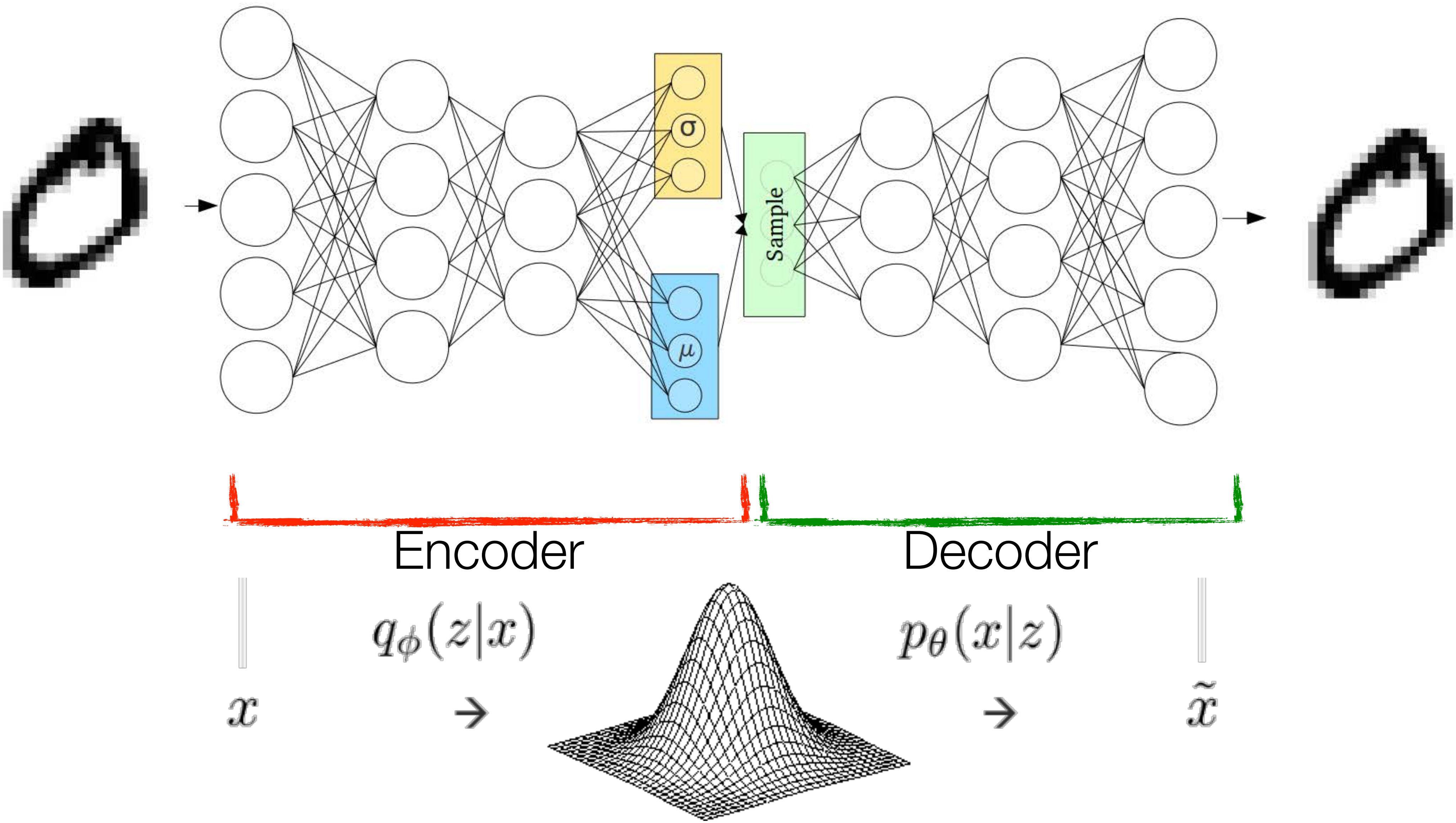
Cobb et al. 2019

- Machine learning approach using Ensemble Neural Nets
- Learns to repeat retrieval of a planet (e.g. WASP-12b) very fast



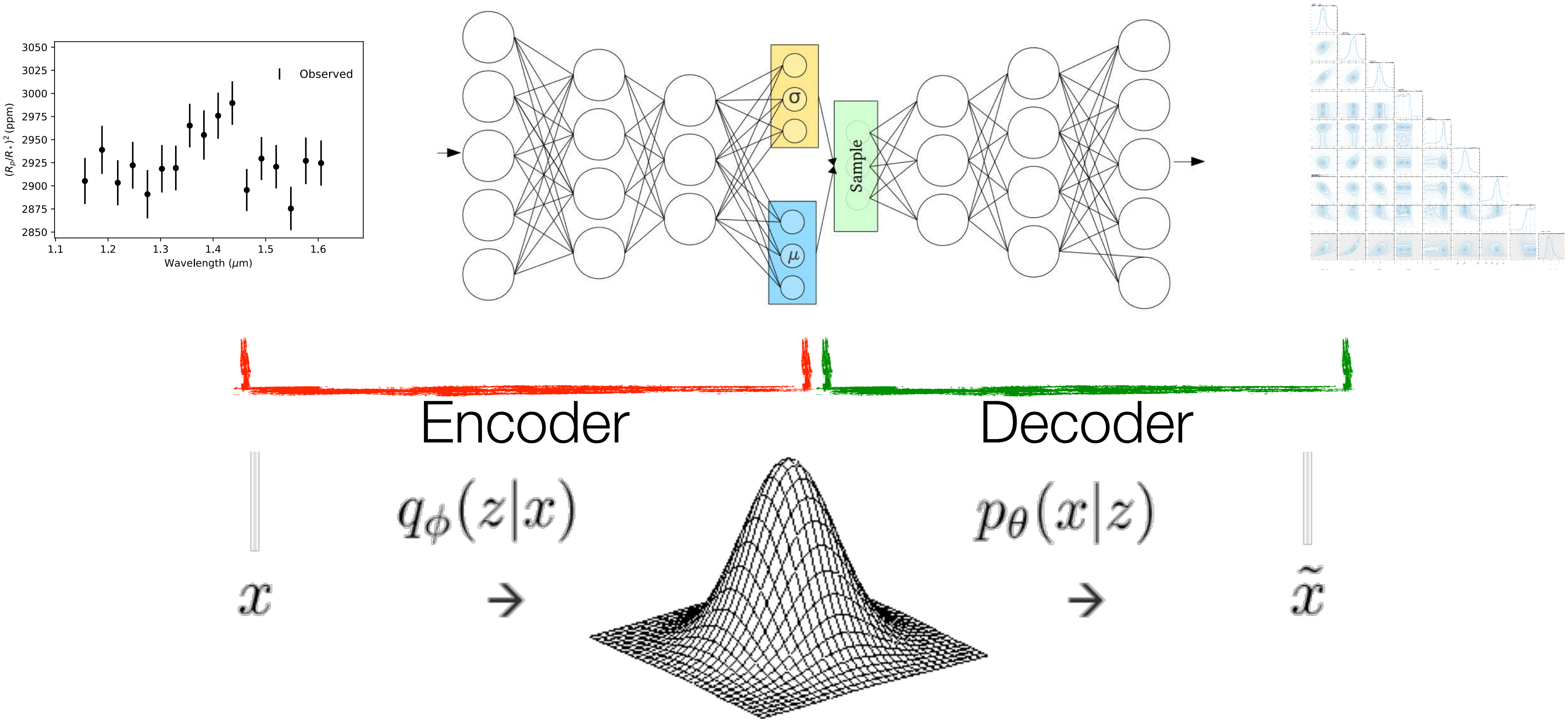
Two schools of thought: to likelihood or not to likelihood

The variational autoencoder



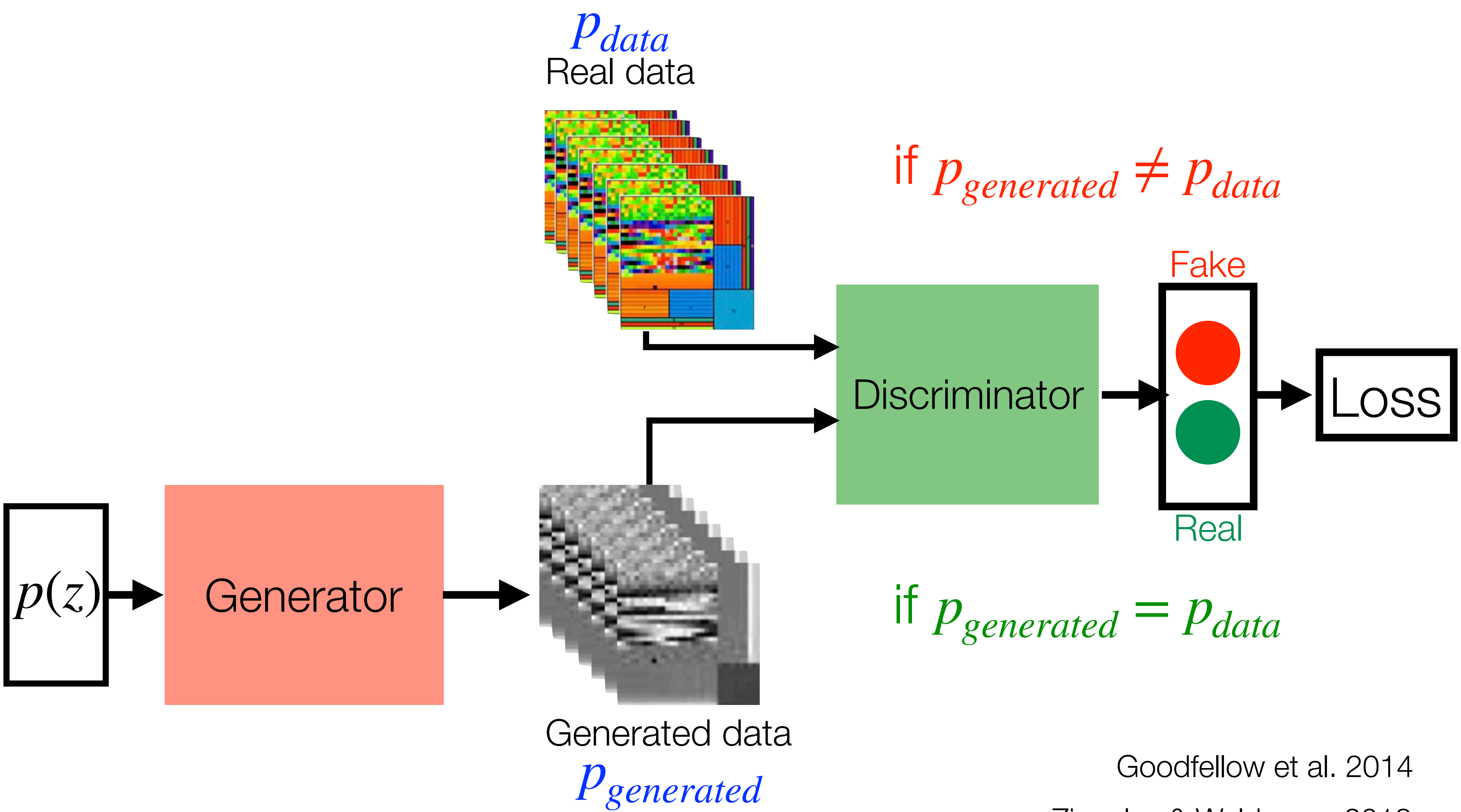
The latent variable space as probabilities

The variational autoencoder



Adversarial Learning: Generative Adversarial Networks

A likelihood free approach



Goodfellow et al. 2014

Zingales & Waldmann 2018

Inpainting

Completing missing information given available data and what the algorithm has learned from training

Missing data
to be completed

Well trained

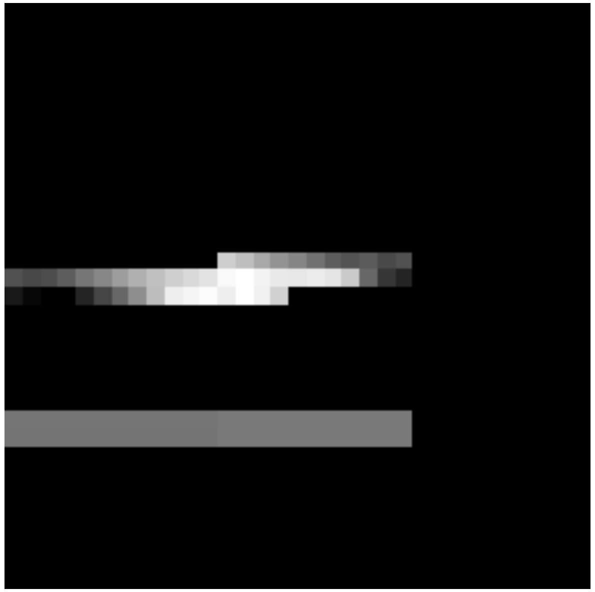
Corrupted

Poorly trained

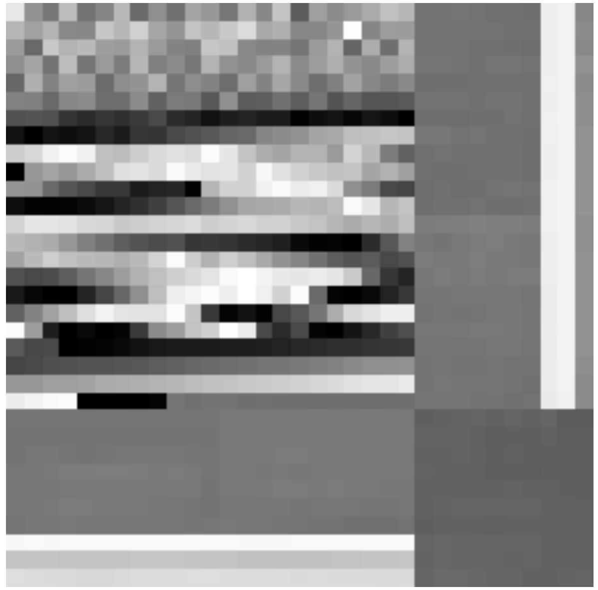
Spectral reconstruction

Spectrum **Parameters**

Input



WFC3 mask

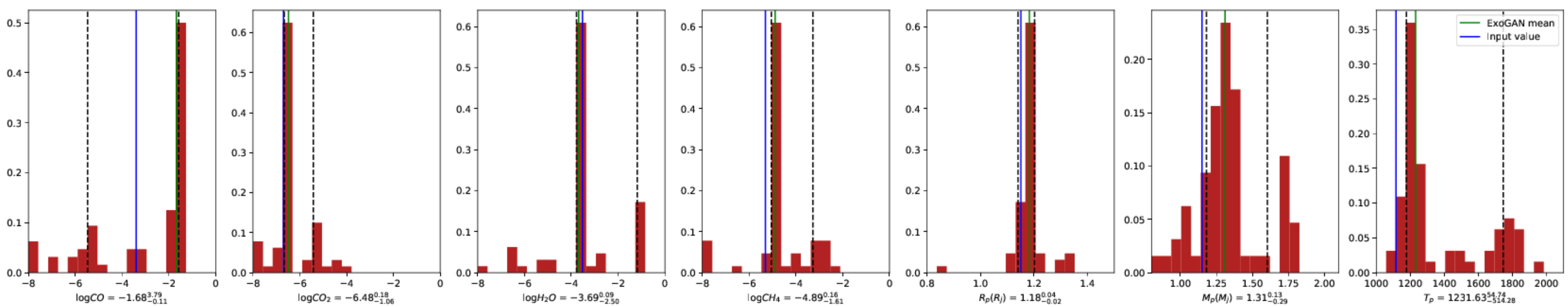


Reconstruction

Zingales & Waldmann 2018

Spectral reconstruction

Test set parameters			
Variable	$A(0\,\sigma_\phi)$	$A(1\,\sigma_\phi)$	$A(2\,\sigma_\phi)$
CO	62.8%	72.6%	78.2%
CO_2	94.2%	96.6%	97.4%
H_2O	89.6%	92.8%	93.9%
CH_4	80.3%	88.2%	91.6%
R_p	100.0%	100.0%	100.0%
M_p	88.0%	89.7%	90.8%
T_p	90.4%	92.2%	93.2%



Zingales & Waldmann 2018

Conclusions

Where to go from here:

- Better understanding and exploiting entropy & sparsity
- Build more tractable neural net architectures
- Formally including data likelihoods

We are hiring!