Machine Learning in Astronomy: An Overview

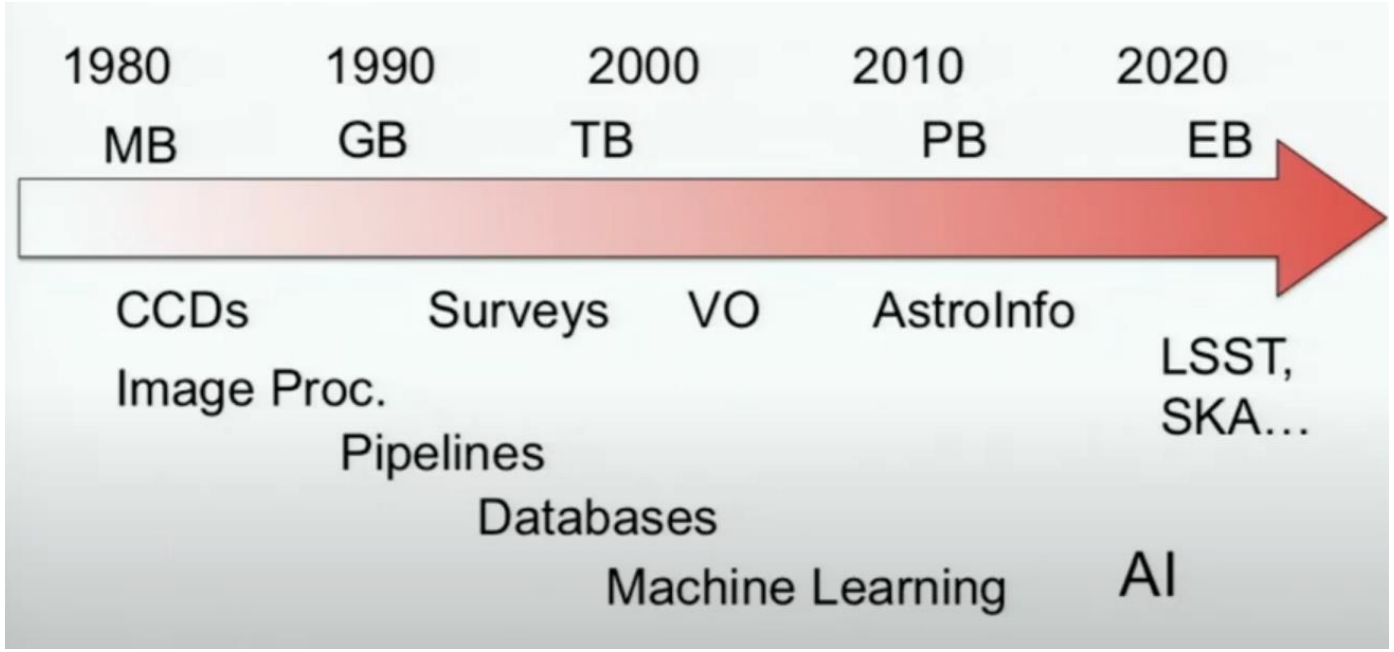
Atul Chhotray [a.chhotray@uva.nl]
Postdoctoral Fellow
Anton Pannekoek Institute (API)

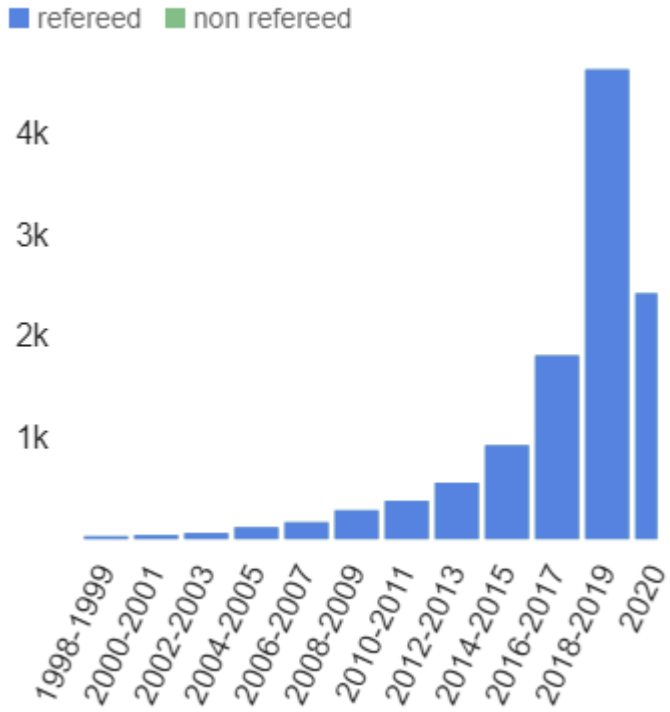
Lecture Overview

- The Big Picture [ML = Machine Learning]
  - Brief History
  - Machine Learning as a Fourth Paradigm
  - Context: Need for ML in Astronomy?

- Some examples of ML in Astronomy
  - Star Galaxy classification
  - Exoplanetary Atmospheres Classification
  - Cosmological Structure Estimation
  - Learning Equations & Launching Black Hole Jets

A Brief History of Machine Learning (in Astronomy)

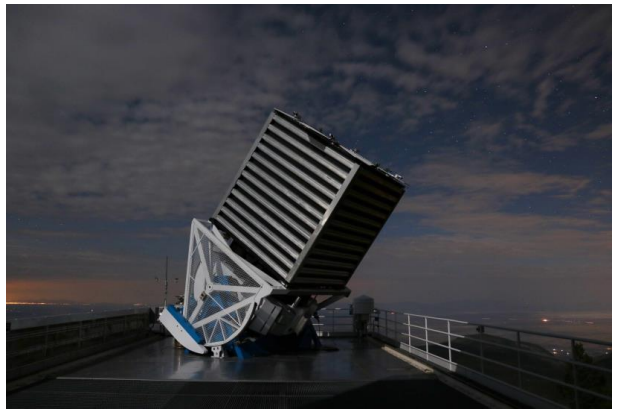




Referred Publications containing the word ‘ML’

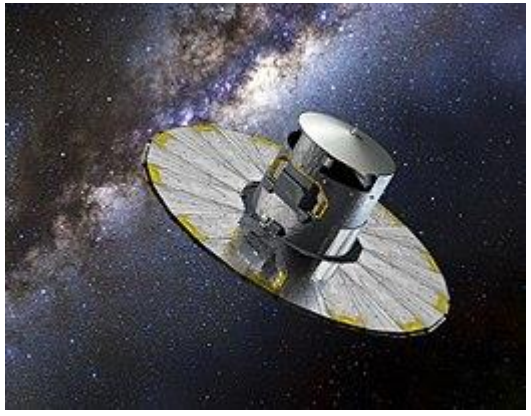
Astronomical Surveys: Googling the Sky!

- Sloan Digital Sky Survey : SDSS
  - 2.5 m telescope
  - Catalogue of > 10^9deep sky objects
  - Generated PB of data!

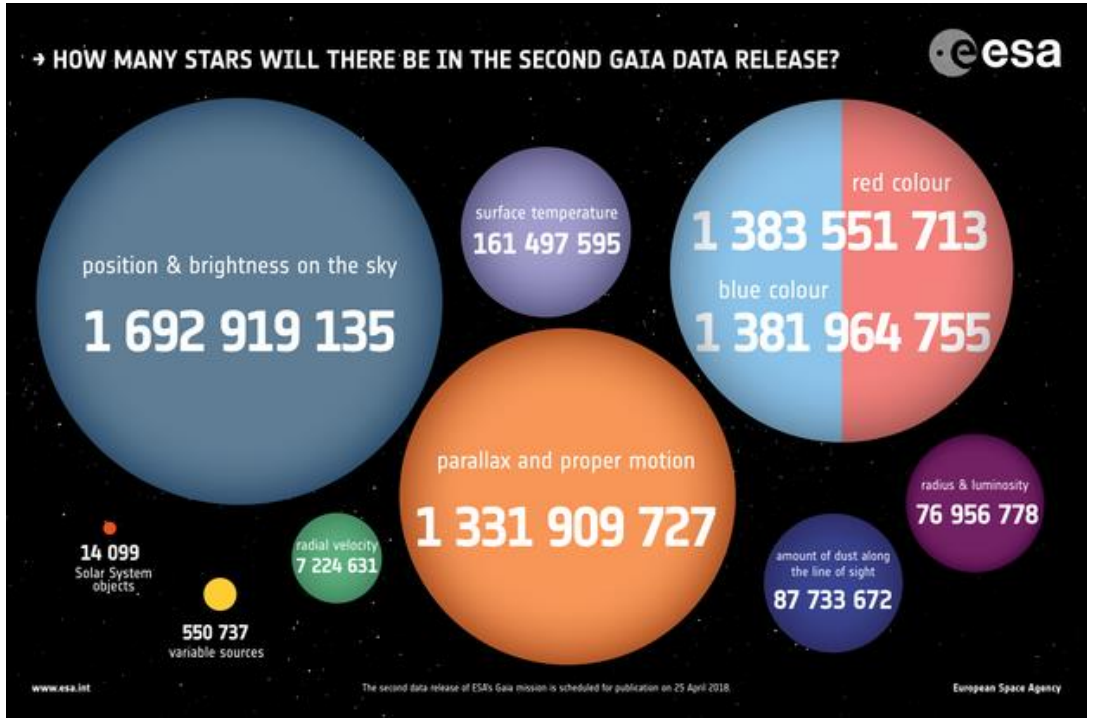


SDSS Telescope. Image Credit: Patrick Gaulme

- Gaia :
  - Precision Astrometry of > 10^9objects
  - Generated PB(?) of data!



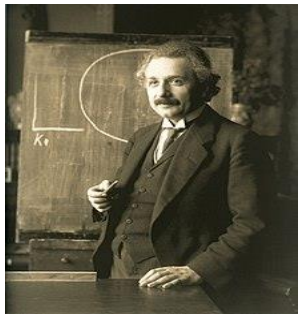
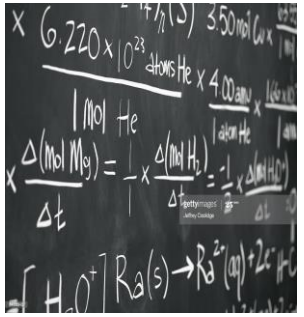
Artistic Impression of Gaia. Image Credit: ESA / ESO

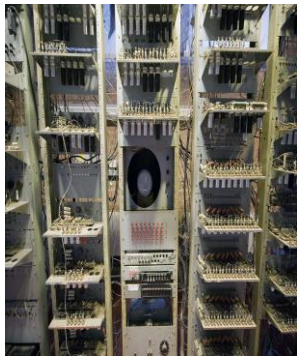


Knowledge Paradigms

- 1st Paradigm: Measurement / Experiment
- 2nd paradigm: Theory / Analytical Work
- 3rd Paradigm : Numerical Simulations



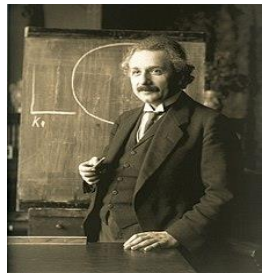
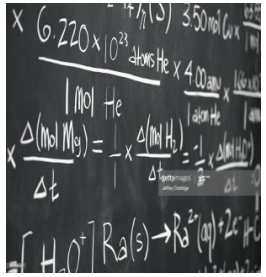



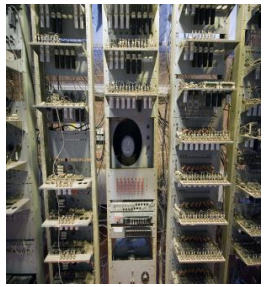
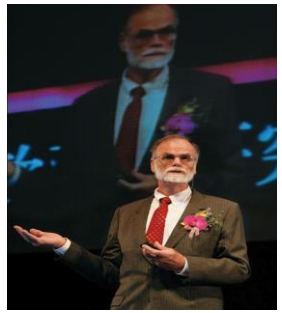



A New Knowledge Paradigm

- 1st Paradigm: Measurement / Experiment
- 2nd paradigm: Theory / Analytical Work
- 3rd Paradigm : Numerical Simulations
- 4th Paradigm : Data Driven Science!

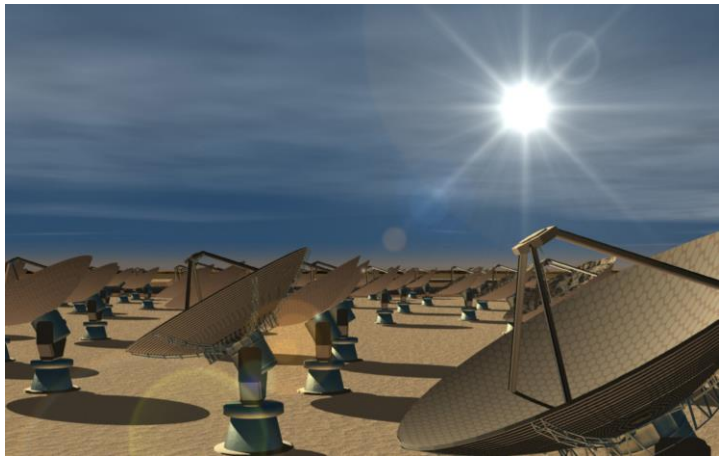



ML in Astronomy : Why?

- Data Deluge in Astronomy
  - Information volume and rate is growing exponentially : Data streams instead of data sets!
  - Existing surveys ~ 100 PB
  - Most data will never be seen by humans
    - Large Synoptic Sky Survey (LSST) ~ 30 TB/night
    - Square Kilometer Array [SKA] (+2022) – 1 EB / sec (raw data)



Artist's impression of SKA:

- Data Information Content is Increasing – Data Driven Science!
  - Multiwavelength information about each object (radio, optical, UV, X-ray….)
  - Gravitational Waves
  - Neutrinos

Artist's impression of LSST:
Credit: lsst.org

- Data Complexity is Increasing – patterns that cannot be comprehended by humans directly
  - Human genome ~ 1 GB
  - 1 TB ~ 2 millions books
  - Human Bandwidth ~ 1 TB / year

http://research.iac.es/winterschool/2018/pages/book-ws2018.php

Star Galaxy Classification

- Latest surveys (LSST, DES etc) collect photometric data for ~$10^9$of astronomical objects (future surveys??)
- Sheer volume makes *Manual* classification impossible
- Use ML to create accurate catalogues
  - Future use
  - Follow up observations (Transients)

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **464,** 4463–4475 (2017) doi:10.1093/mnras/stw2672
Advance Access publication 2016 October 17

Star–galaxy classification using deep convolutional neural networks

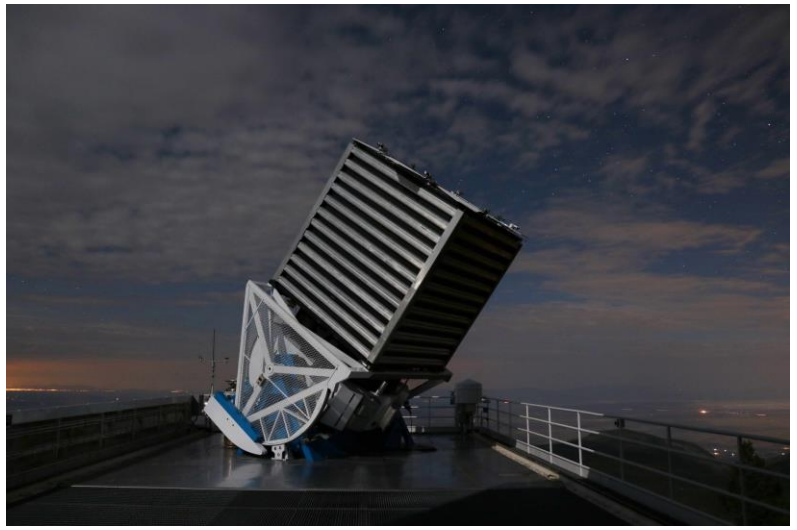
Edward J. Kim[1]★ and Robert J. Brunner[1,2,3,4]

[1]*Department of Physics, University of Illinois, Urbana, IL 61801, USA*
[2]*Department of Astronomy, University of Illinois, Urbana, IL 61801, USA*
[3]*Department of Statistics, University of Illinois, Champaign, IL 61820, USA*
[4]*National Center for Supercomputing Applications, Urbana, IL 61801, USA*

Accepted 2016 October 12. Received 2016 October 4; in original form 2016 June 25

Star Galaxy Classification

- Data:
  - Sloan Digital Sky Survey (SDSS): >300 millions objects
    - Randomly select 65000 Sources = 47656 galaxies + 17344 stars
  - Canada France Hawaii Telescope Lensing Survey (CFHTLenS) > 25 million objects
    - Selection of Galaxies = 57843, Stars = 8545
  - Image size after extraction: 48 x 48 pixels



SDSS Telescope
Image Credit: Patrick Gaulme

CFHT Telescope
Image Credit: cfht.Hawaii.edu

~~Star Galaxy Classification~~

- Deep Convolutional Neural Network
  - Also used in computer vision community

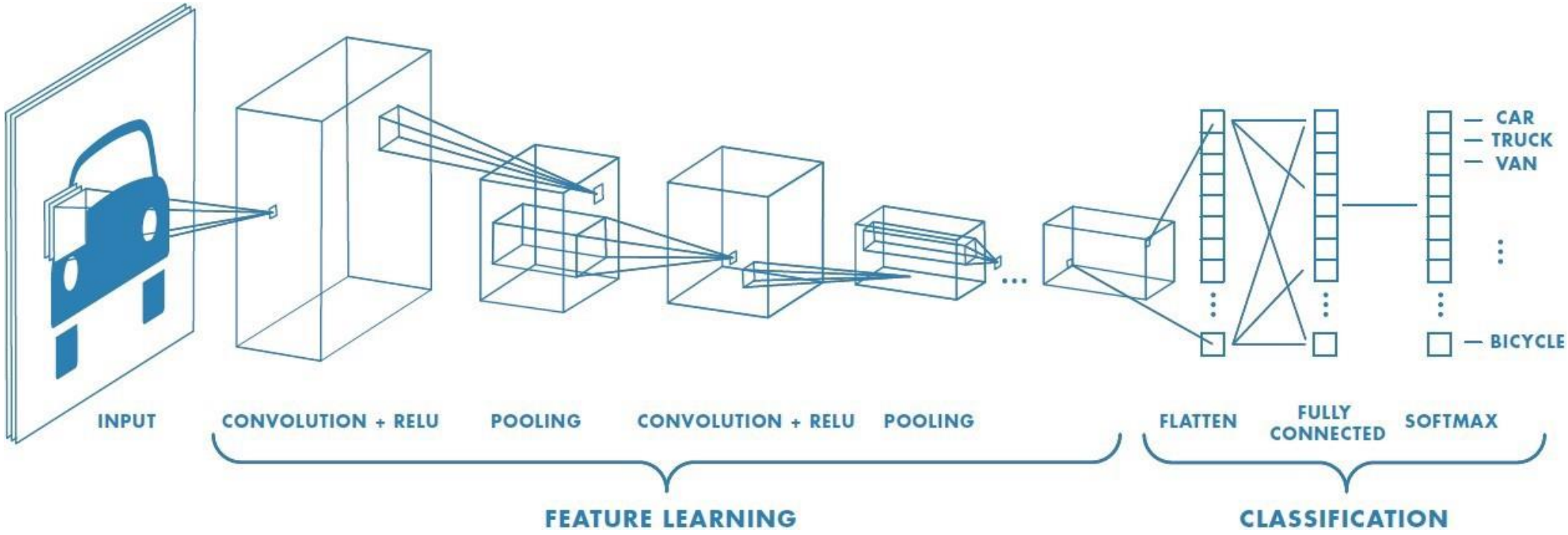


Image Credit: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Star Galaxy Classification

- **ConvNet**
- 11 layers
- Image in 5 bands (ugriz)
- First layer has 32 filters of size 5 x 5 x 5
- Max Pooling : 2 x 2
- Fully connected layers in the end

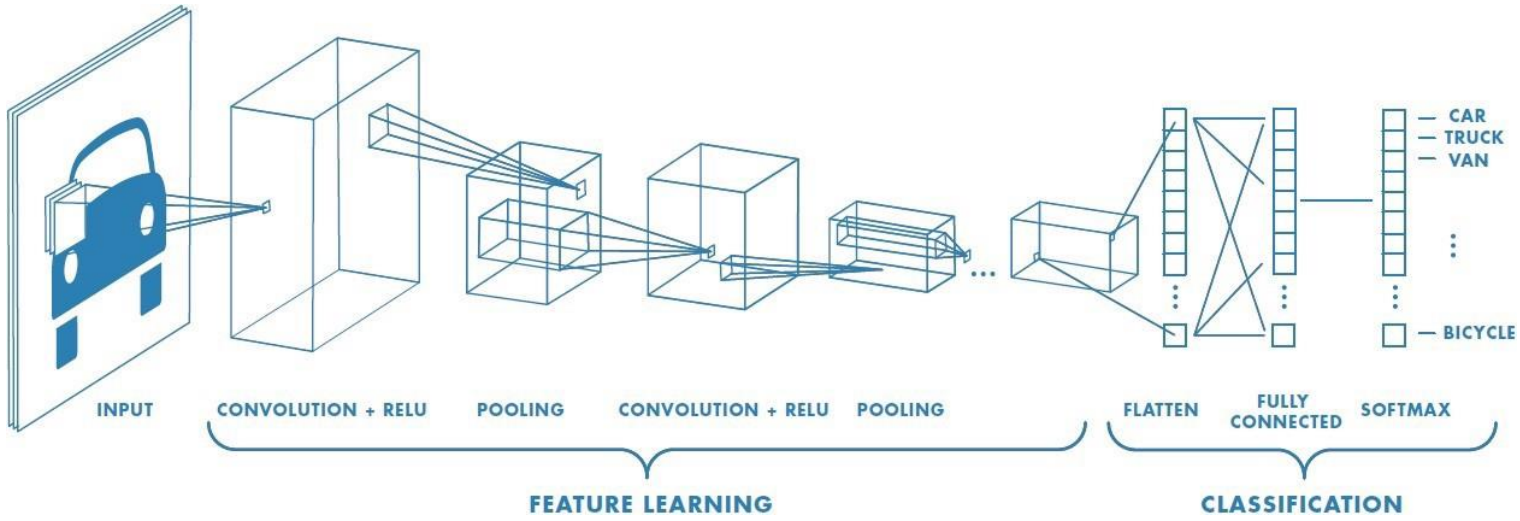


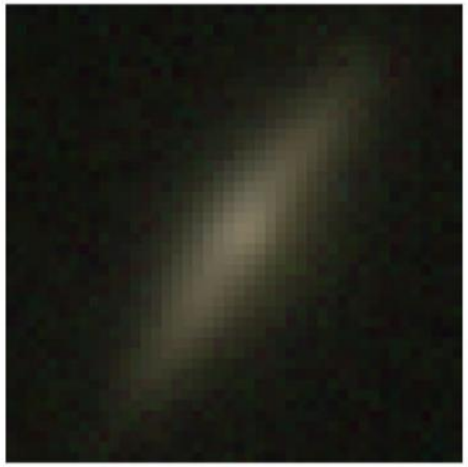
Table 1. Summary of ConvNet architecture and hyperparameters. Note that pooling layers have no learnable parameters.

Type	Filters	Filter size	Padding	Non-linearity	Initial weights	Initial biases
Convolutional	32	5×5	–	Leaky ReLU	Orthogonal	0.1
Convolutional	32	3×3	1	Leaky ReLU	Orthogonal	0.1
Pooling	–	2×2	–	–	–	–
Convolutional	64	3×3	1	Leaky ReLU	Orthogonal	0.1
Convolutional	64	3×3	1	Leaky ReLU	Orthogonal	0.1
Convolutional	64	3×3	1	Leaky ReLU	Orthogonal	0.1
Pooling	–	2×2	–	–	–	–
Convolutional	128	3×3	1	Leaky ReLU	Orthogonal	0.1
Convolutional	128	3×3	1	Leaky ReLU	Orthogonal	0.1
Convolutional	128	3×3	1	Leaky ReLU	Orthogonal	0.1
Pooling	–	2×2	–	–	–	–
Fully connected	2048	–	–	Leaky ReLU	Orthogonal	0.01
Fully connected	2048	–	–	Leaky ReLU	Orthogonal	0.01
Fully connected	2	–	–	Softmax	Orthogonal	0.01

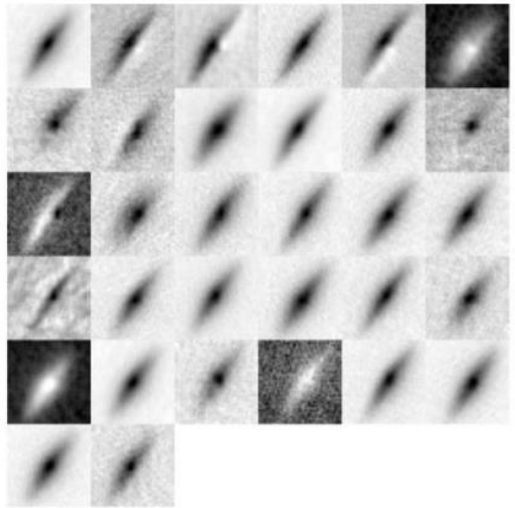
Star Galaxy Classification

- Learnable parameters : 11 million
- Training Set: < million … we have a problem

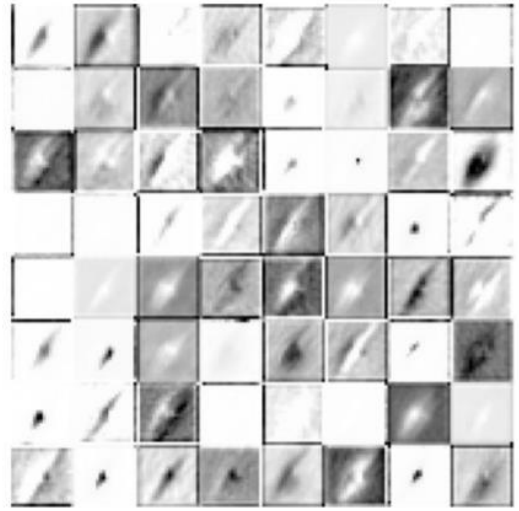
- Dealing with Overfitting
  - Data Augmentation : Transform each image via label preserving transformation
    - Rotation, Reflection, Translation, Gaussian Noise
  - Dropout : Randomly setting the output of each hidden neuron in previous layer to zero [helps learn more robustly]
  - Bayesian Model Combination : Generate ensemble combinations of models



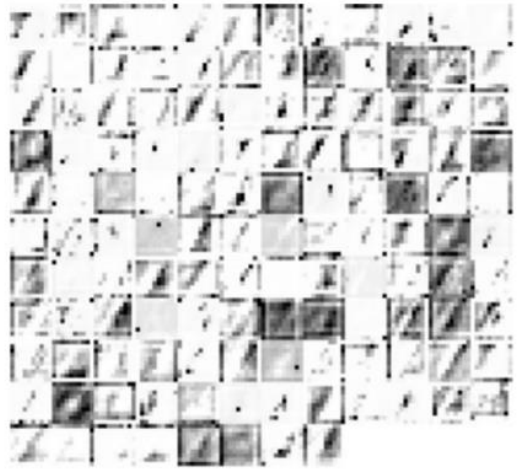
(a) Input (5 bands×44×44)



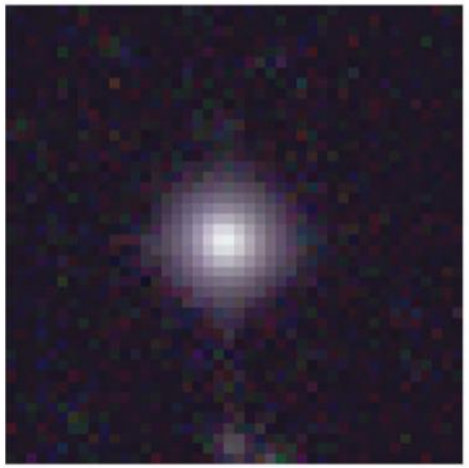
(b) Layer 1 (32 maps×40×40)



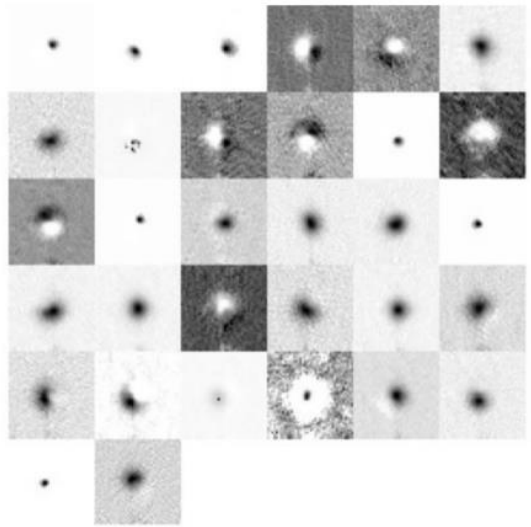
(c) Layer 3 (64 maps×20×20)



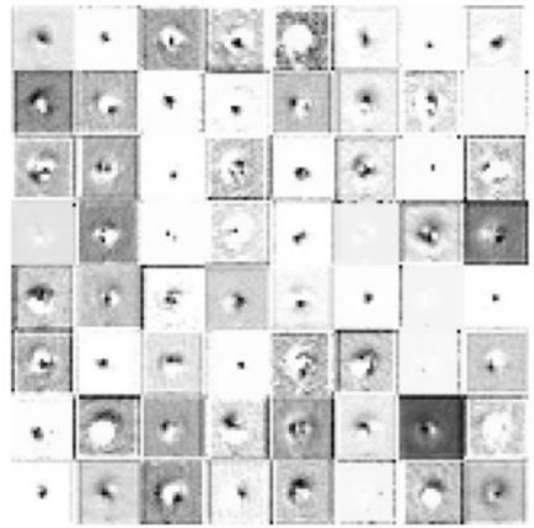
(d) Layer 6 (128 maps×10×10)



(a) Input (5 bands×44×44)



(b) Layer 1 (32 maps×40×40)



(c) Layer 3 (64 maps×20×20)

(d) Layer 6 (128 maps×10×10)

Star Galaxy Classification

- Metrics and Model Comparison
  - Several metrics used

- Compare with Decision Trees and Random Forest: TPC algorithm
  - TPCphot – fewer attributes
  - TPCmorph – more attributes, including handcrafted feature

Table 2. The definition of the classification performance metrics.

Metric	Meaning
AUC	Area under the receiver operating curve
MSE	Mean squared error
c_g	Galaxy completeness
p_g	Galaxy purity
c_s	Star completeness
p_s	Star purity
$p_g(c_g = x)$	Galaxy purity at x galaxy completeness
$c_s(p_s = x)$	Star completeness at x star purity
CAL	Calibration error with overlapping binning
$\lvert\Delta N_g\rvert/N_g$	Absolute error in number of galaxies
log loss	Cross-entropy

$$p_g = \frac{N_g}{N_g + M_s} \qquad c_g = \frac{N_g}{N_g + M_g}$$

Star Galaxy Classification

- For CFHTLenS dataset:

Classifier	AUC	MSE	$p_g(c_g = 0.96)$	$c_s(p_s = 0.97)$	CAL	$\lvert\Delta N_g\rvert/N_g$	log loss
ConvNet	**0.9948**	0.0112	**0.9972**	0.8971	**0.0197**	**0.0029**	**0.0441**
$\text{TPC}_{\text{morph}}$	0.9924	**0.0109**	0.9963	**0.9268**	0.0245	0.0056	0.0809
TPC_{phot}	0.9876	0.0189	0.9927	0.8044	0.0266	0.0101	0.1085

- For SDSS dataset:

Classifier	AUC	MSE	$p_g(c_g = 0.96)$	$c_s(p_s = 0.97)$	CAL	$\lvert\Delta N_g\rvert/N_g$	log loss
ConvNet	0.9952	0.0182	0.9915	0.9500	**0.0243**	0.0157	**0.0731**
$\text{TPC}_{\text{morph}}$	**0.9967**	**0.0099**	**0.9977**	**0.9810**	0.0254	**0.0044**	0.0914
TPC_{phot}	0.9886	0.0283	0.9819	0.8879	0.0316	0.0160	0.1372

Other Classification Algorithms

- Stellar Classification
- Star / Galaxy / Quasar
- Variable Stars
- Galaxy Morphology Prediction

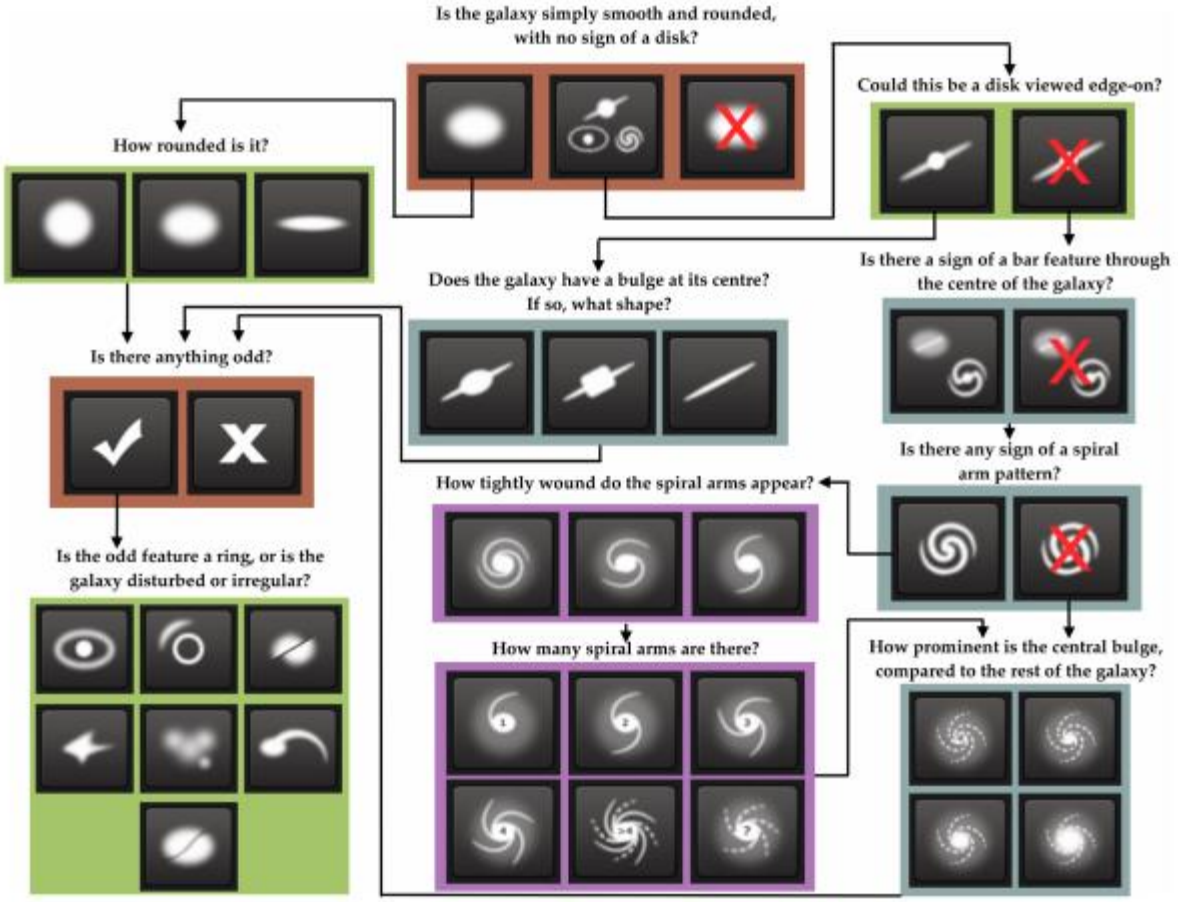


Figure 1. The Galaxy Zoo 2 decision tree. Reproduced from fig. 1 in Willett et al. (2013).

Exoplanets

- Did you hear News about Venus these past weeks??

Exoplanets

- News about Venus??
- Detection of Phosphene gas – a possible biomarker
- Identifying gases in exoplanet atmospheres!

nature astronomy

ARTICLES

https://doi.org/10.1038/s41550-020-1174-4

Check for updates

Phosphine gas in the cloud decks of Venus

Jane S. Greaves[1,2]✉, Anita M. S. Richards[3], William Bains[4], Paul B. Rimmer[5,6,7], Hideo Sagawa[8], David L. Clements[9], Sara Seager[4,13,14], Janusz J. Petkowski[4], Clara Sousa-Silva[4], Sukrit Ranjan[4], Emily Drabek-Maunder[1,10], Helen J. Fraser[11], Annabel Cartwright[1], Ingo Mueller-Wodarg[9], Zhuchang Zhan[4], Per Friberg[12], Iain Coulson[12], E'lisa Lee[12] and Jim Hoge[12]

Measurements of trace gases in planetary atmospheres help us explore chemical conditions different to those on Earth. Our nearest neighbour, Venus, has cloud decks that are temperate but hyperacidic. Here we report the apparent presence of phosphine (PH_3) gas in Venus's atmosphere, where any phosphorus should be in oxidized forms. Single-line millimetre-waveband spectral detections (quality up to ~15σ) from the JCMT and ALMA telescopes have no other plausible identification. Atmospheric PH_3 at ~20 ppb abundance is inferred. The presence of PH_3 is unexplained after exhaustive study of steady-state chemistry and photochemical pathways, with no currently known abiotic production routes in Venus's atmosphere, clouds, surface and subsurface, or from lightning, volcanic or meteoritic delivery. PH_3 could originate from unknown photochemistry or geochemistry, or, by analogy with biological production of PH_3 on Earth, from the presence of life. Other PH_3 spectral features should be sought, while in situ cloud and surface sampling could examine sources of this gas.

BBC Sign in News Sport Reel Worklife Travel Future

NEWS

Home US Election Coronavirus Video World UK Business Tech Science

Science & Environment

Is there life floating in the clouds of Venus?

By Jonathan Amos
BBC Science Correspondent

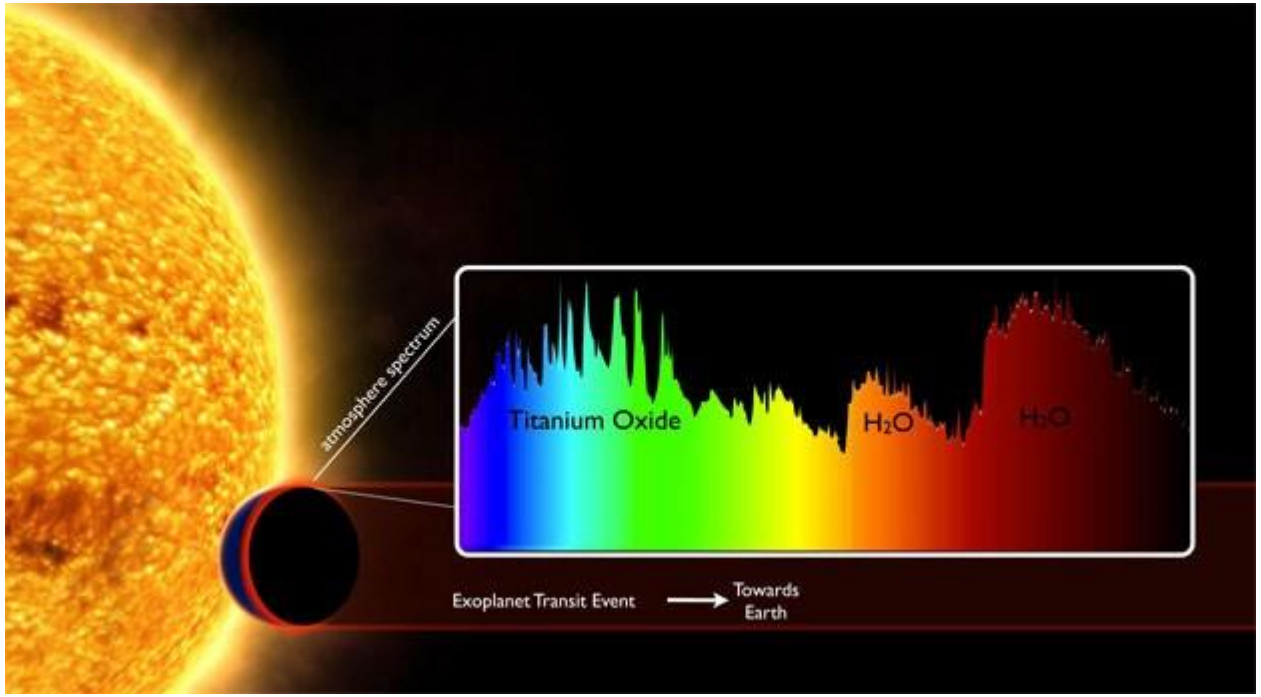
14 September 2020 Share

The New York Times

Life on Venus? Astronomers See a Signal in Its Clouds

The detection of a gas in the planet's atmosphere could turn scientists' gaze to a planet long overlooked in the search for extraterrestrial life.

Classifying Exoplanet Emission Spectra



DREAMING OF ATMOSPHERES

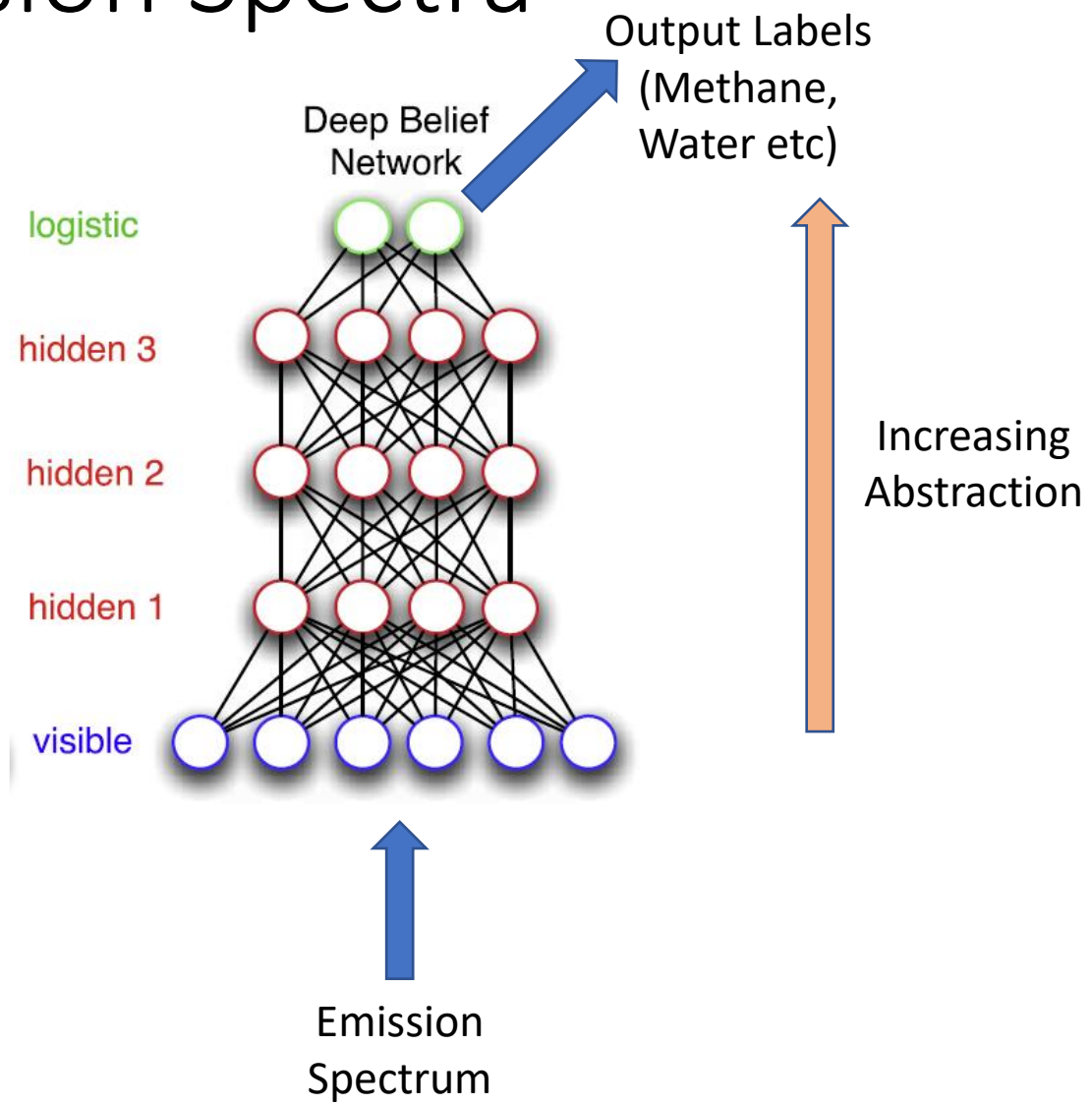
I. P. WALDMANN
Department of Physics & Astronomy, University College London, Gower Street, WC1E 6BT, UK; ingo@star.ucl.ac.uk
Received 2015 November 20; accepted 2016 February 10; published 2016 March 28

ABSTRACT

Here, we introduce the *RobERt* (Robotic Exoplanet Recognition) algorithm for the classification of exoplanetary emission spectra. Spectral retrieval of exoplanetary atmospheres frequently requires the preselection of molecular/

Classifying Exoplanet Emission Spectra

- Deep Belief Neural Network
  - Mimics human recognition of spectra
  - Hidden layers form an abstract representation of underlying features
- Training
  - Supply a input spectrum with a label
    - 5 planets
    - 17150 spectra / planet
    - No mixtures considered
  - Mini Batch Stochastic Gradient Descent
  - Use Dropout algorithms



Classifying Exoplanet Emission Spectra

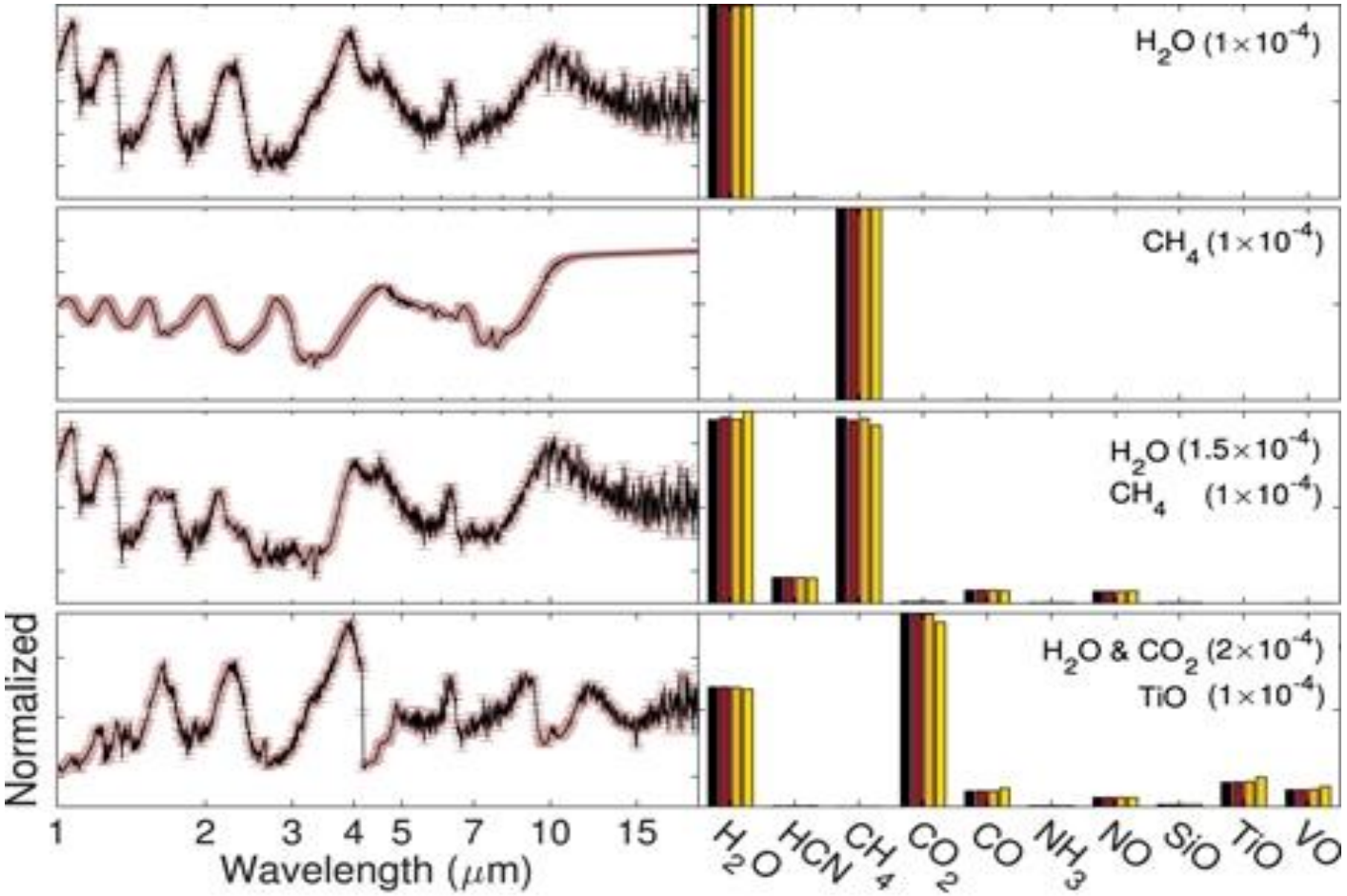


Figure 3 from Dreaming of Atmospheres
I. P. Waldmann 2016 ApJ 820 107 doi:10.3847/0004-637X/820/2/107

Dreaming of Atmospheres

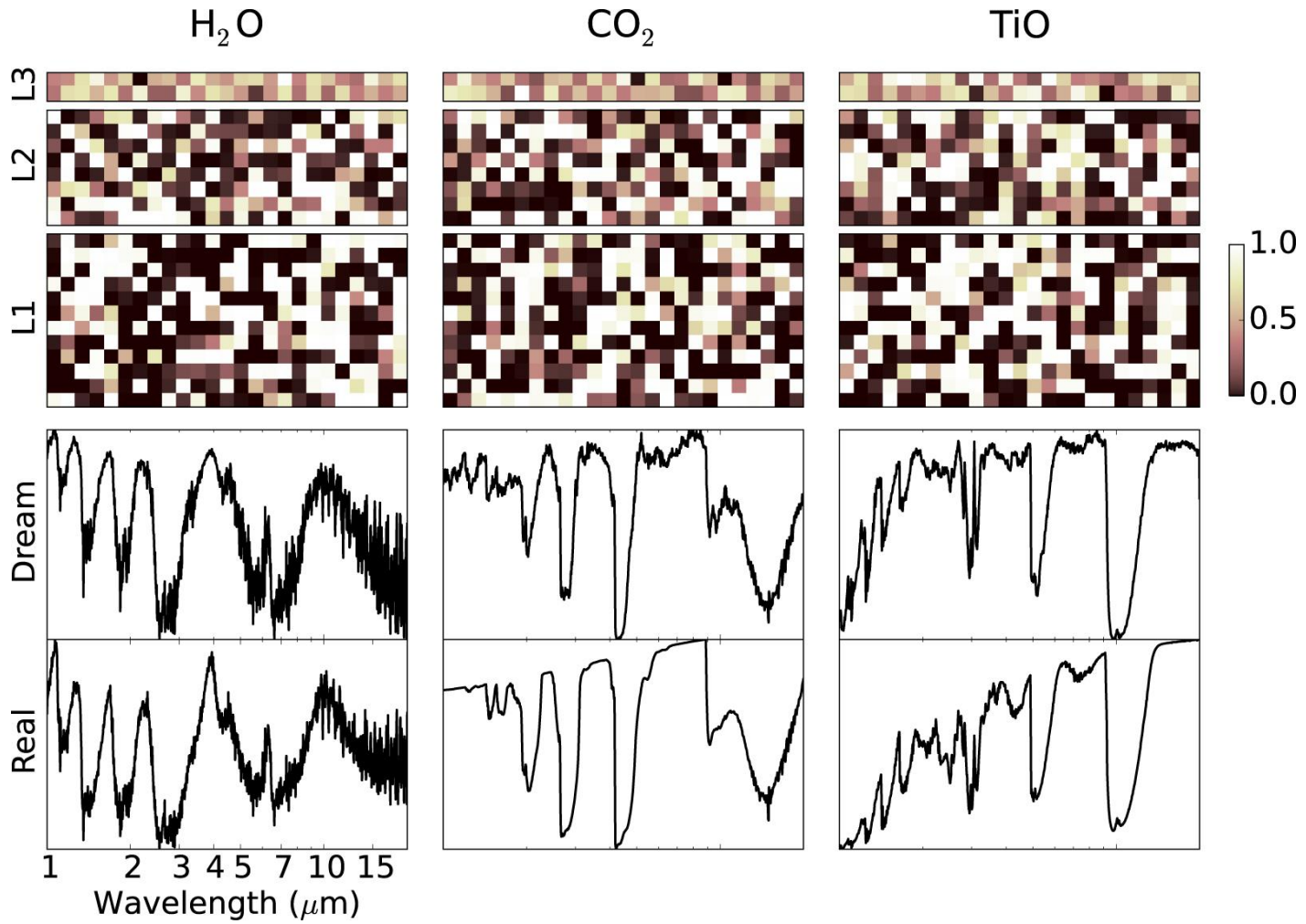


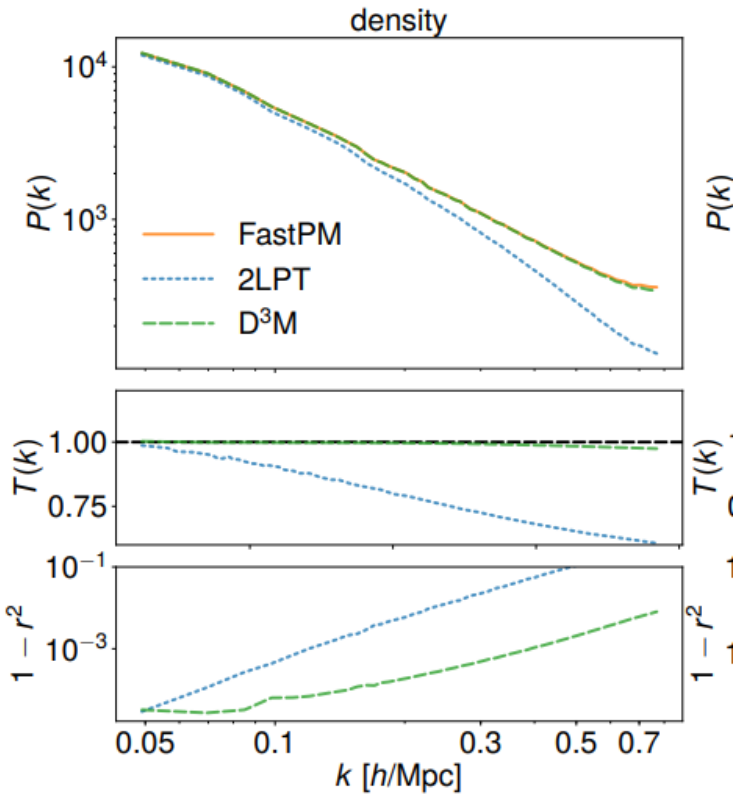
Figure 5. Spectral reconstruction (or "dreaming") of three molecules H_2O, CO_2, and TiO. The top three panels show neuron activations for the bottom (L1) to top (L3) hidden layers. The bottom two rows show normalized H_2O, CO_2, and TiO spectra reconstructed by the neural network and real data examples as comparison. The similarities between the "dreamed" and real spectral features are striking. This indicates a good representation of molecular features in the neural network.

ML and Simulations : Data Driven Simulations

- "Learning to Predict the Cosmological Structure Formation"
  - Previous work!
    - Compare with survey observations
    - Run N body simulations (computationally very extensive)
  - Use Deep Neural Network to predict large scale cosmological structure
    - Trained using N body cosmological simulation data set
    - Compare with semi analytical approx

Learning to Predict the Cosmological Structure Formation

Siyu He[a,b,c,1], Yin Li[d,e,f], Yu Feng[d,e], Shirley Ho[c,e,d,a,b,1], Siamak Ravanbakhsh[g], Wei Chen[c], and Barnabás Póczos[h]



ML and Semi Analytical Theory

- Launching jets from Black Holes
  - GR (General Relativity)
  - MHD (Magnetohydrodynamics)
  - Extremely complex set of equations to solve to find jet solutions
  - Multi-dimensional parameter space extremely time consuming to explore semi-analytically
  - Use ML to predict solutions!
    - ANN
    - SVM

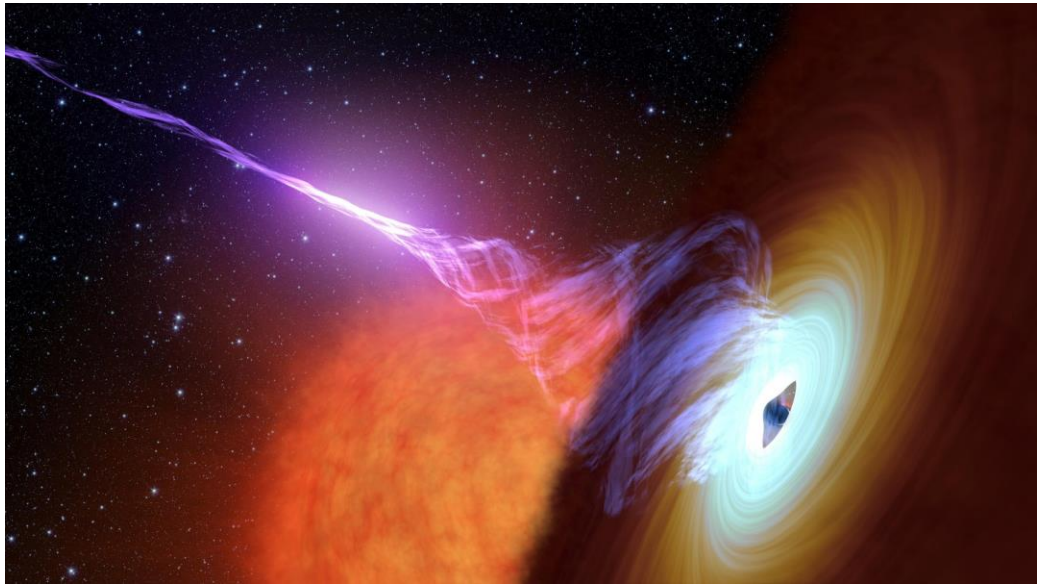


Figure: Black hole launching jet

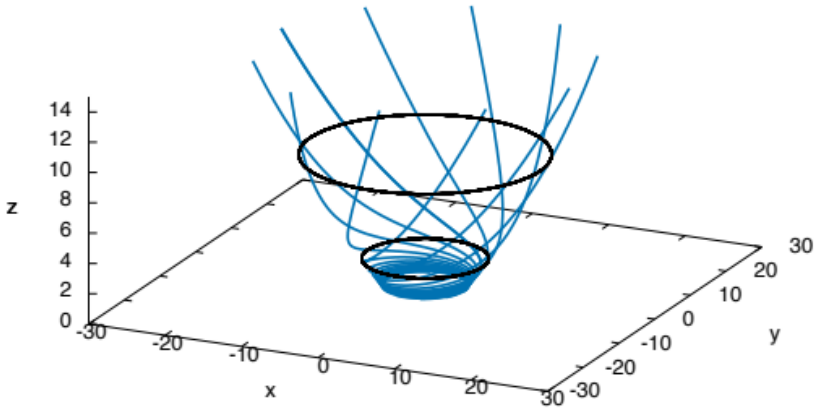


Figure: Magnetic Field lines for a solution