Applications of Machine Learning in Astronomy

Alexander Chaushev

Project Scientist @ UCI / Visitor @ UCSD
(a.chaushev@uci.edu)

San Diego Machine Learning Meetup - April 5th

Image courtesy of *NASA/FUSE/Lynette Cook*

Talk Structure

1. Astronomy (~15 minutes)
   - Key areas of astronomical research
   - What do we know about exoplanets?
   - How hard is it to find an exoplanet?

2. Machine learning application (~15 minutes)
   - Transit surveys for exoplanets
   - Key problem
   - Deep neural networks for candidate classification

1. Astronomy Overview

Alex's Top 3 Mysteries of the Universe!*

There are more planets than stars (100 billion+) in our galaxy.

Is there life?

$10^{\sim 4\text{-}5}$ km

Cosmic Distance Scale

*reasons for an existential crisis

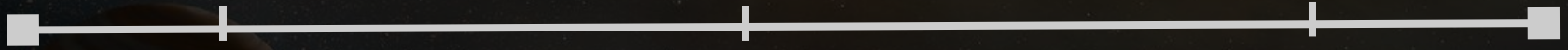
Alex's Top 3 Mysteries of the Universe!*

There are more planets than stars (100 billion+) in our galaxy.

Is there life?

Dark matter influences the structure of galaxies and is 5 x more than regular matter.

What is it?



$10^{\sim 4\text{-}5}$ km

$10^{\sim 18}$ km

Cosmic Distance Scale

*reasons for an existential crisis

Alex's Top 3 Mysteries of the Universe!*

There are more planets than stars (100 billion+) in our galaxy.

Is there life?

$10^{\sim 4\text{-}5}$ km

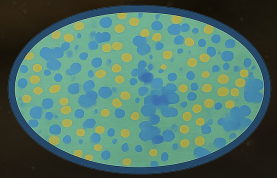
Dark matter influences the structure of galaxies and is 5 x more than regular matter.

What is it?

$10^{\sim 18}$ km

Dark energy makes up 73% of energy in the universe causing it to expand.

Where does it come from?



$10^{\sim 24}$ km

Cosmic Distance Scale

*reasons for an existential crisis

Alex's Top 3 Mysteries of the Universe!*

There are more planets than stars (100 billion+) in our galaxy.

Is there life?

$10^{\sim 4\text{-}5}$ km

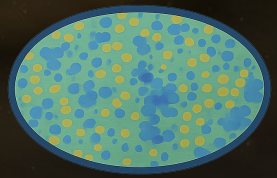
Dark matter influences the structure of galaxies and is 5 x more than regular matter.

What is it?

$10^{\sim 18}$ km

Dark energy makes up 73% of energy in the universe causing it to expand.

Where does it come from?

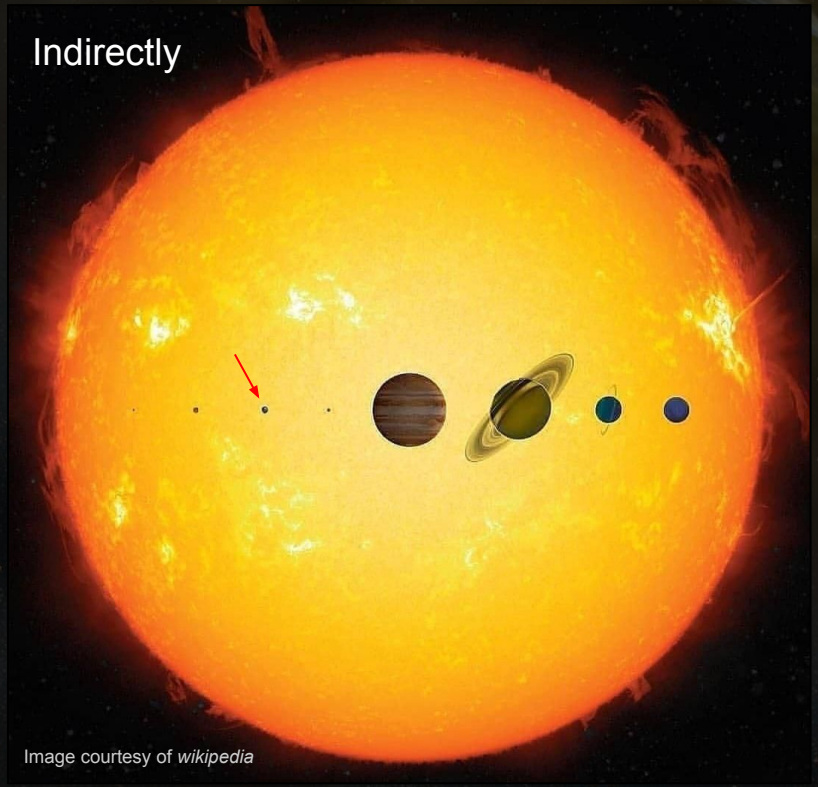


$10^{\sim 24}$ km

Cosmic Distance Scale

*reasons for an existential crisis

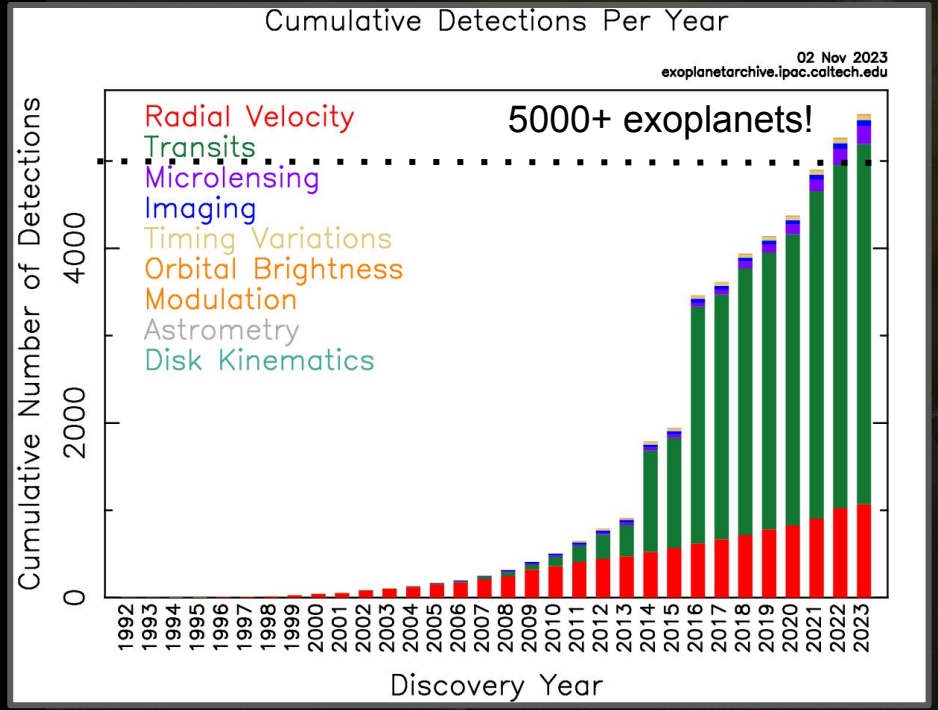
Planets are hard to spot…

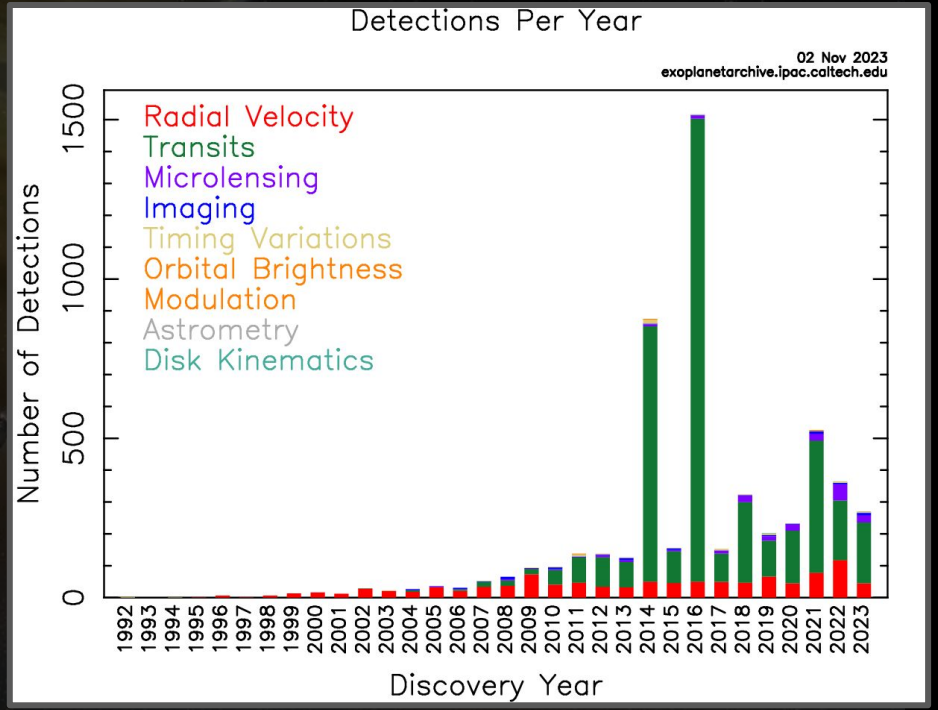


- Voyager 1's 'Pale Blue Dot' taken 6 billion kilometers from Earth.

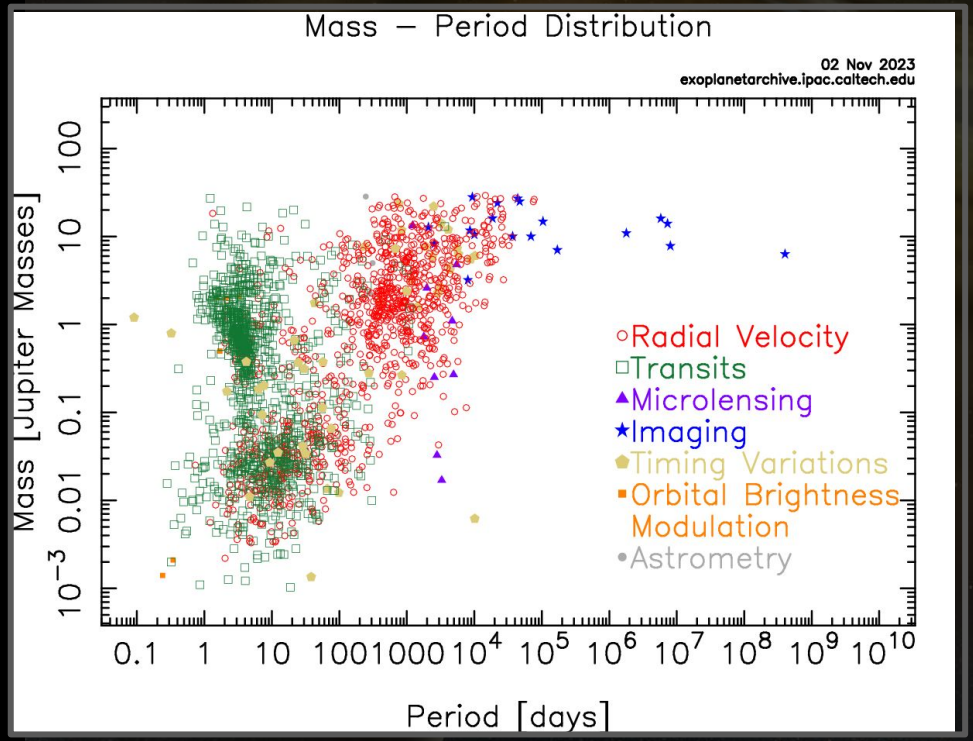
- The nearest star (Proxima Centauri) is 40 trillion kilometers away or ~6,700 times further

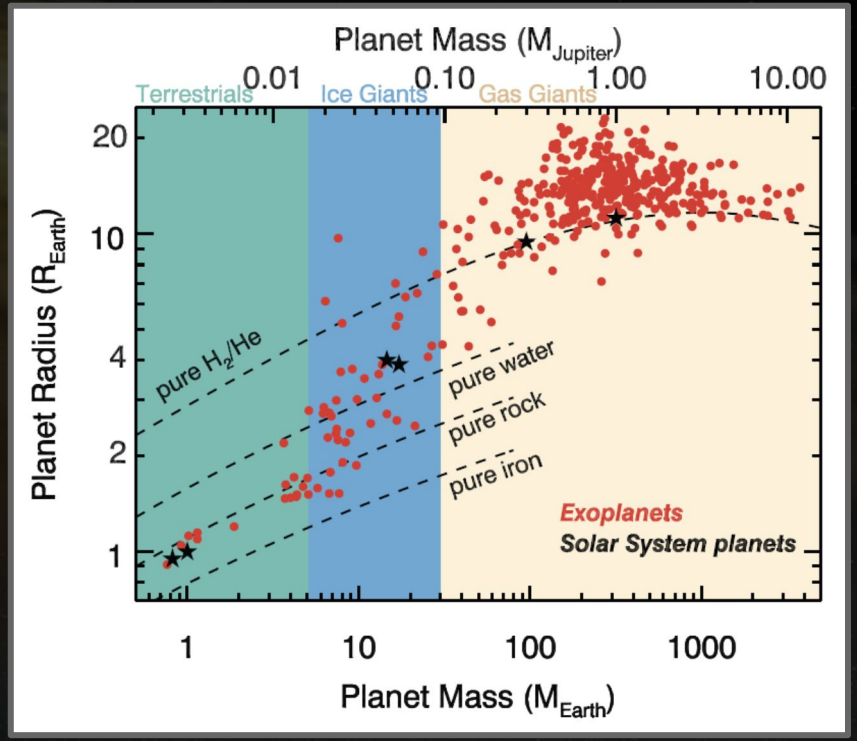
Number of discovered exoplanets





What do these exoplanets look like?

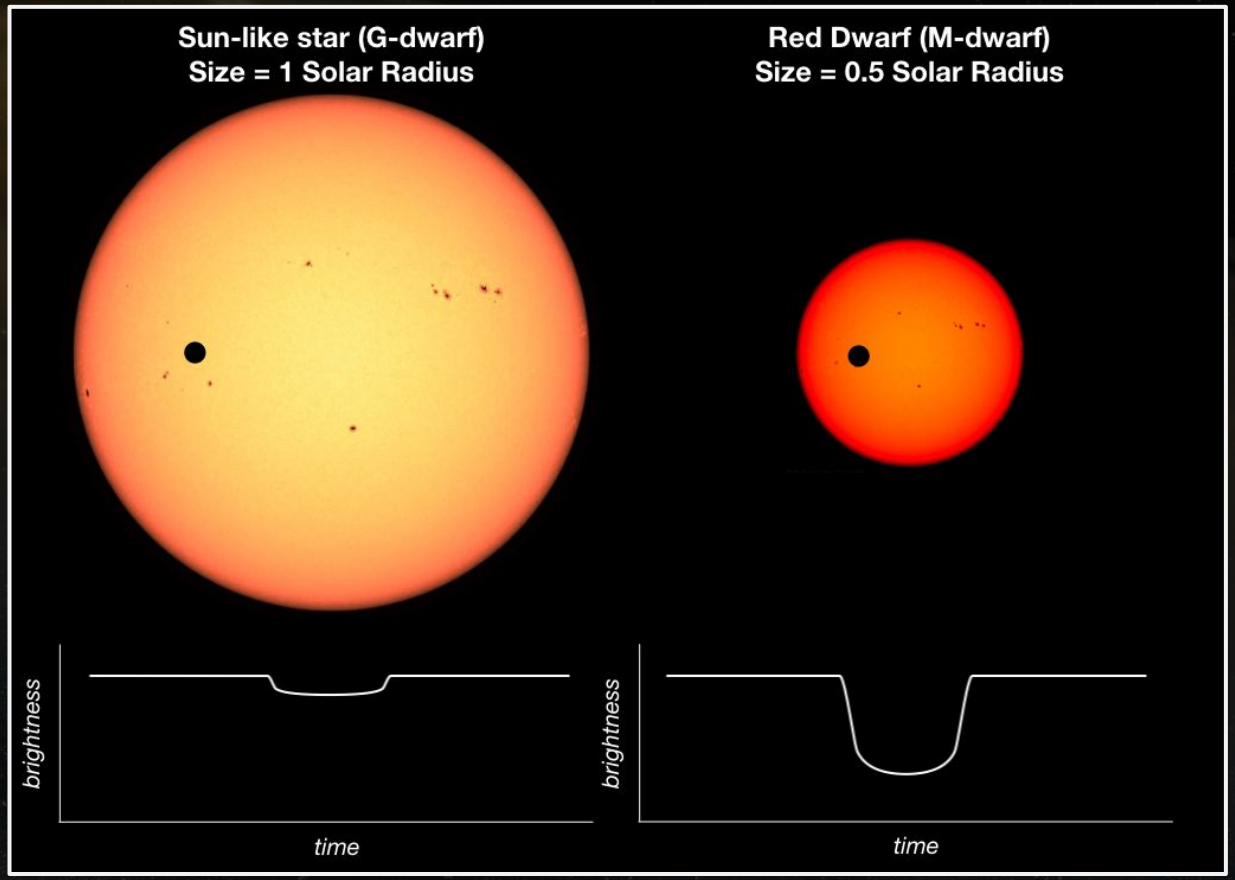




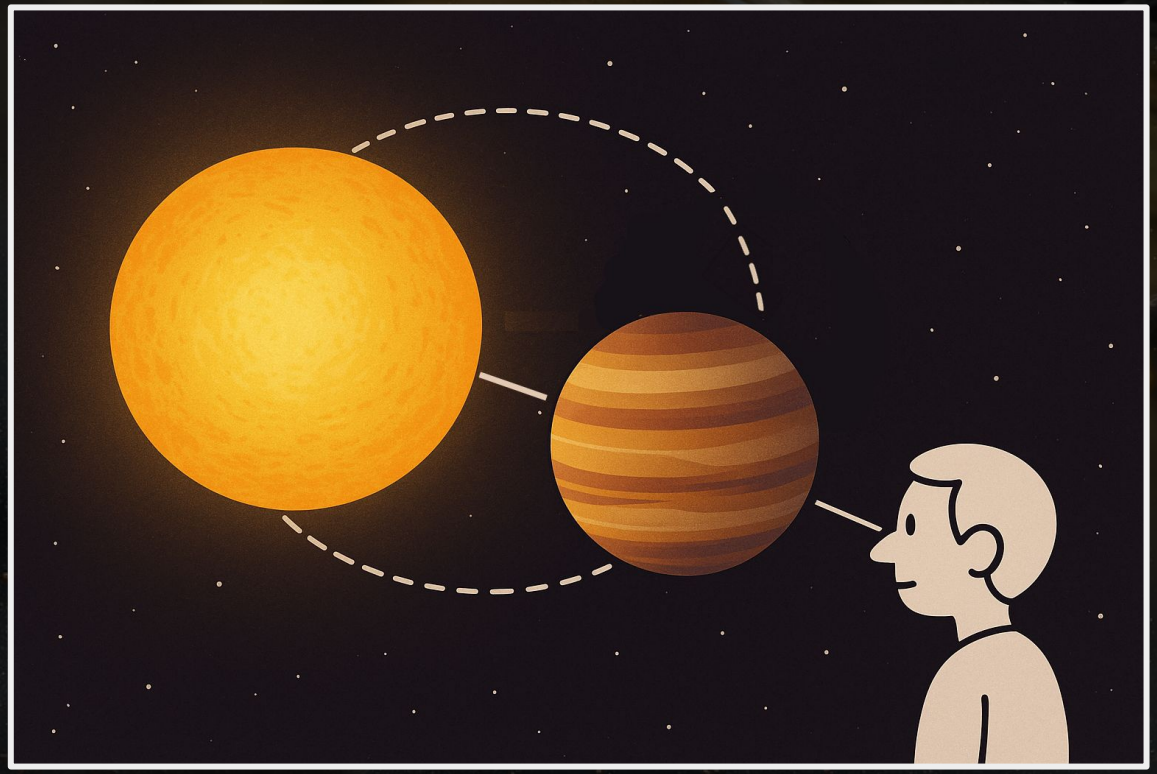
The transit method

- Vast majority are found indirectly through effects on star (radial velocity + transits)

- Transit depth in lightcurve is the ratio of areas of planet and star:
  - Jupiter ~ 1%
  - Neptune ~ 0.1%
  - Earth ~ 0.01%



The transit method



Transit probability equation:

$$P_{transit} = \sim R_s/a$$

a: distance of the planet from the star

Rs: the star's radius.

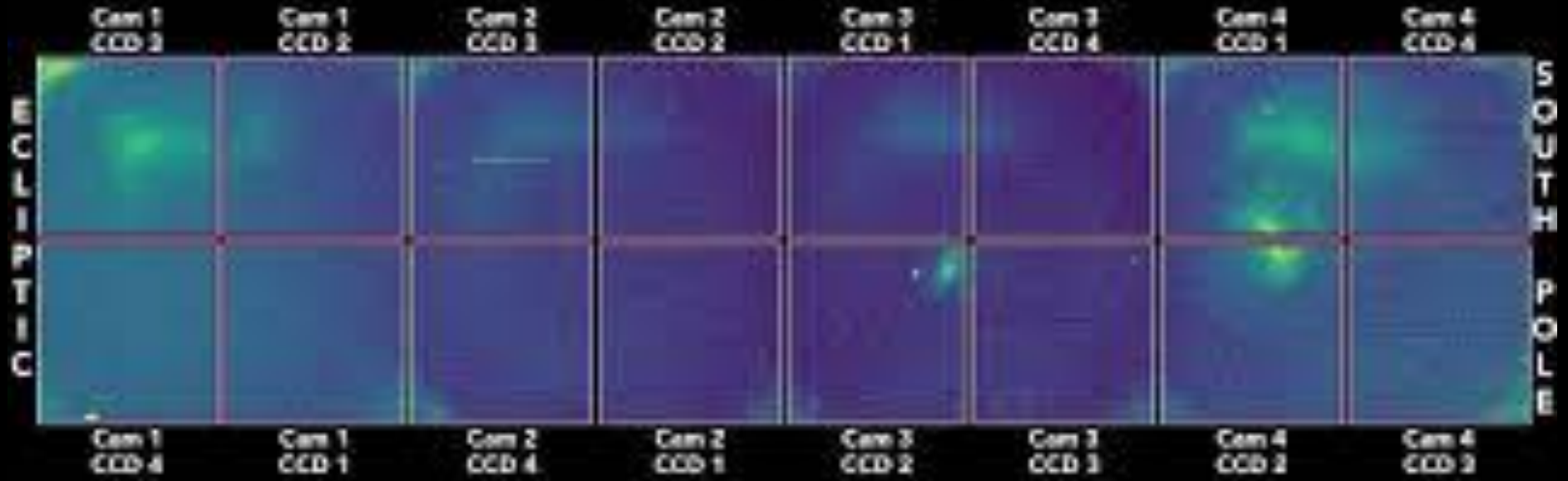
For example:

- Transit probability of Earth at 1 au is 0.5%
- Earth at 0.01 AU is 5%

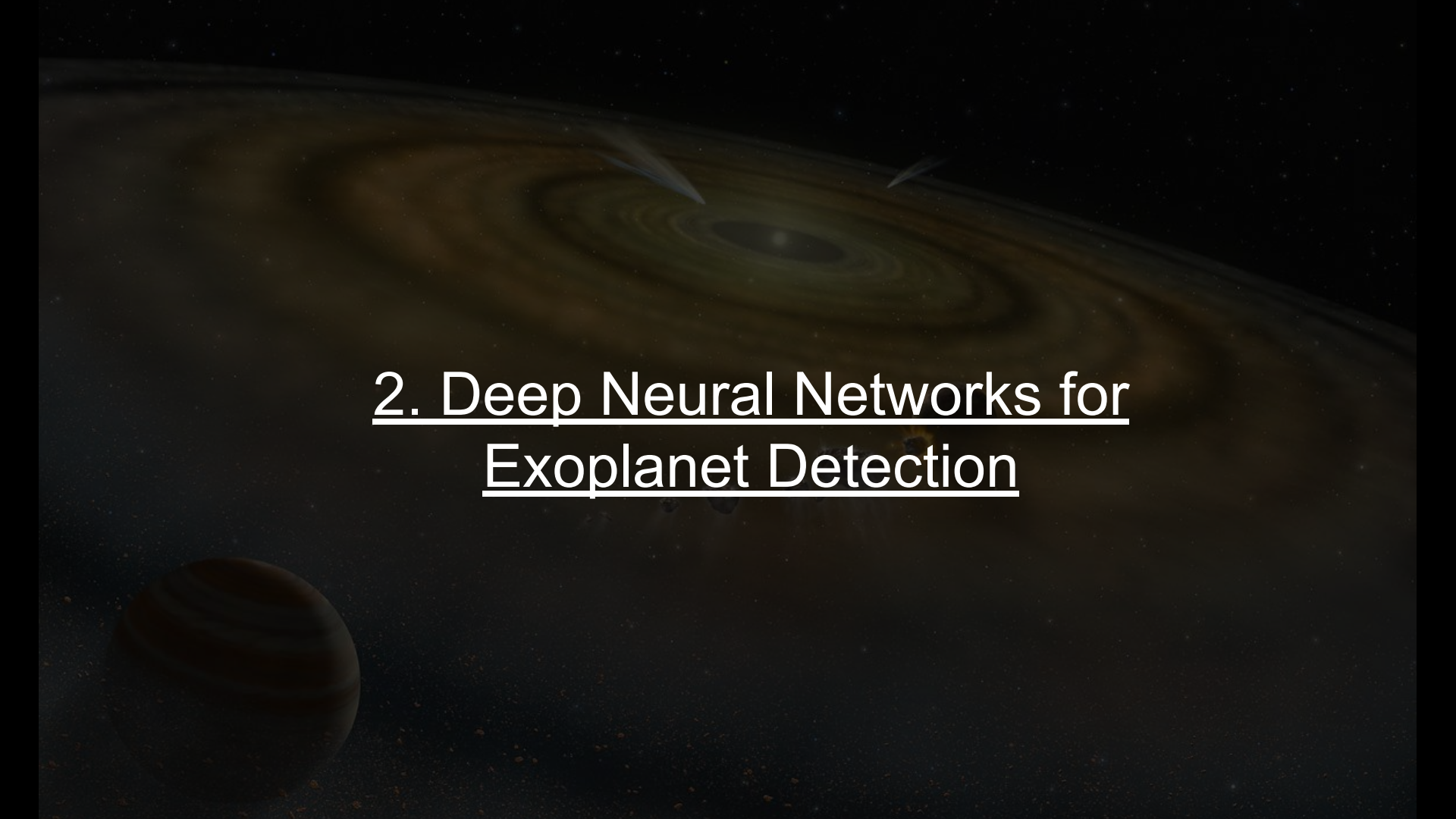
TESS: The Movie

Sector 1

09 Aug 2018 19:44



By Ethan Kruse
@ethan_kruse



2. Deep Neural Networks for Exoplanet Detection

Example Transit Surveys

The Next Generation Transit Survey (NGTS)

- Operational since 2017
- Based in ESO Paranal, Chile at 2,440m altitude
- Found 21 planets
- 12 ground-based 20cm telescopes

Example dataset:
110 billion measurements of 600,000 stars as of July 2022

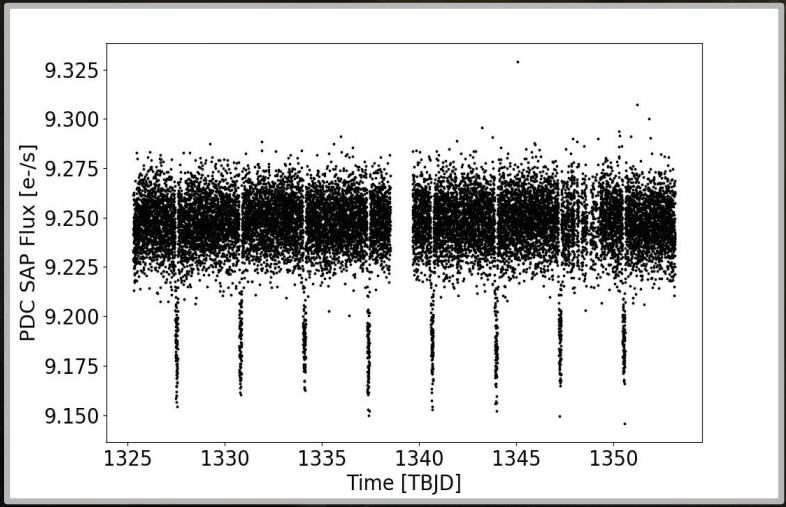
Transiting Exoplanet Survey Satellite (TESS)

- Operational since 2018
- Found 620 planets with 7,525 candidates still under investigation
- Based in space, on a 2:1 lunar resonant orbit called P/2
-

Example dataset (1 sector of 100+):
2,000-10,000 stars a high cadence (<2 min) + images every 5-10 minutes for 25 days

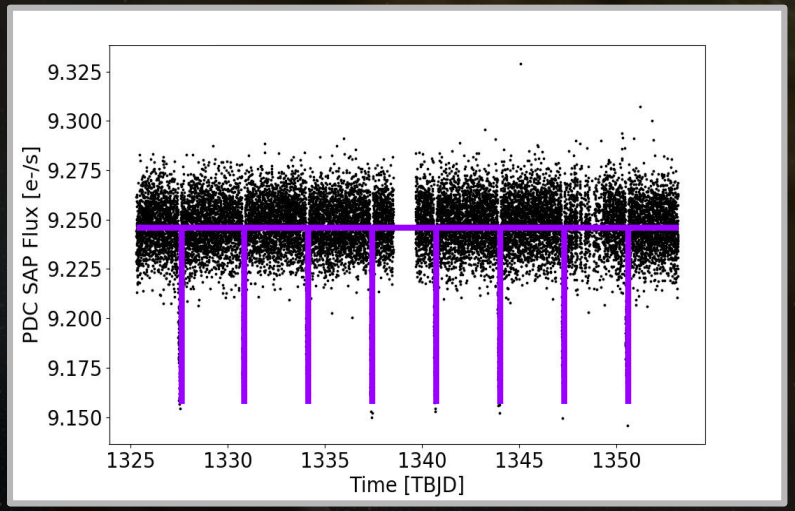
Detecting an exoplanet transit

The TESS lightcurve of WASP-126 b



Detecting an exoplanet transit

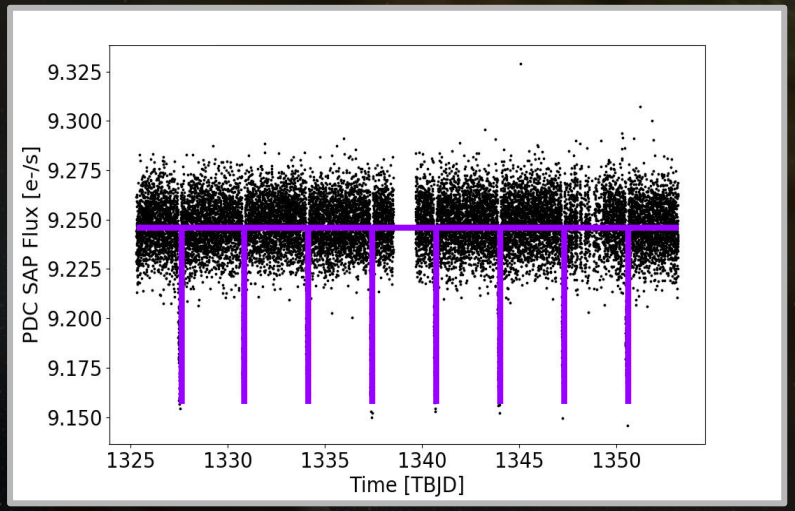
The TESS lightcurve of WASP-126 b



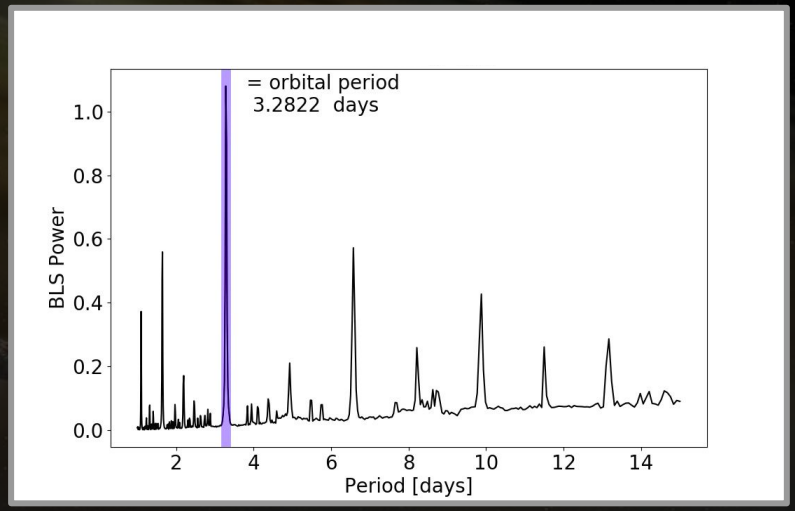
BLS finds candidates which are then vetted by experts

Detecting an exoplanet transit

The TESS lightcurve of WASP-126 b



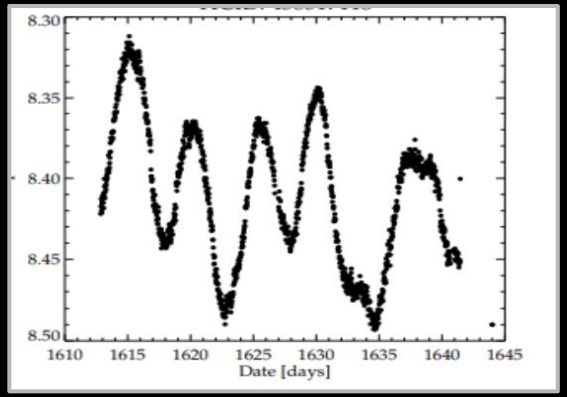
Box-least squares periodogram (Kovacs, 2002)



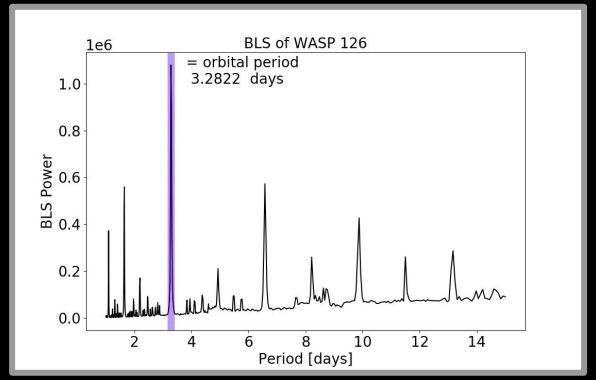
BLS finds candidates which are then vetted by experts

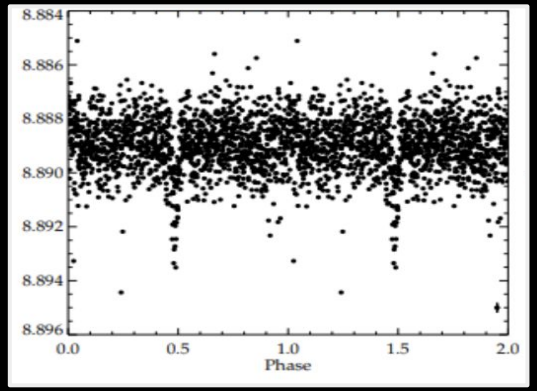
Many sources of false-positives exist

Box-least squares periodogram (Kovacs, 2002)

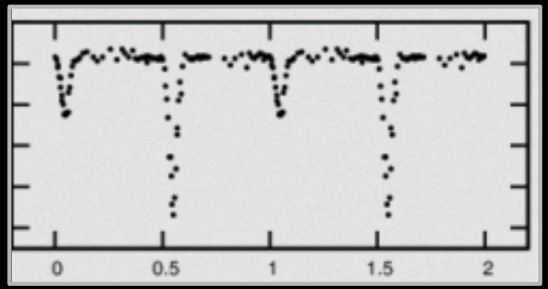


A highly variable young star





Transit or instrument systematic?

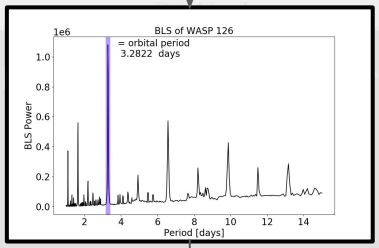


Stellar binary?

Figures adapted from Oelkers (2018)

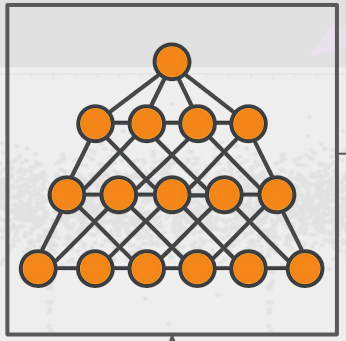
Exoplanet Detection Pipeline

Data: 1,000s of astronomical time series



10,000s of candidate signals

DNN Classifier



Planet

Not planet

Network Architecture + Data

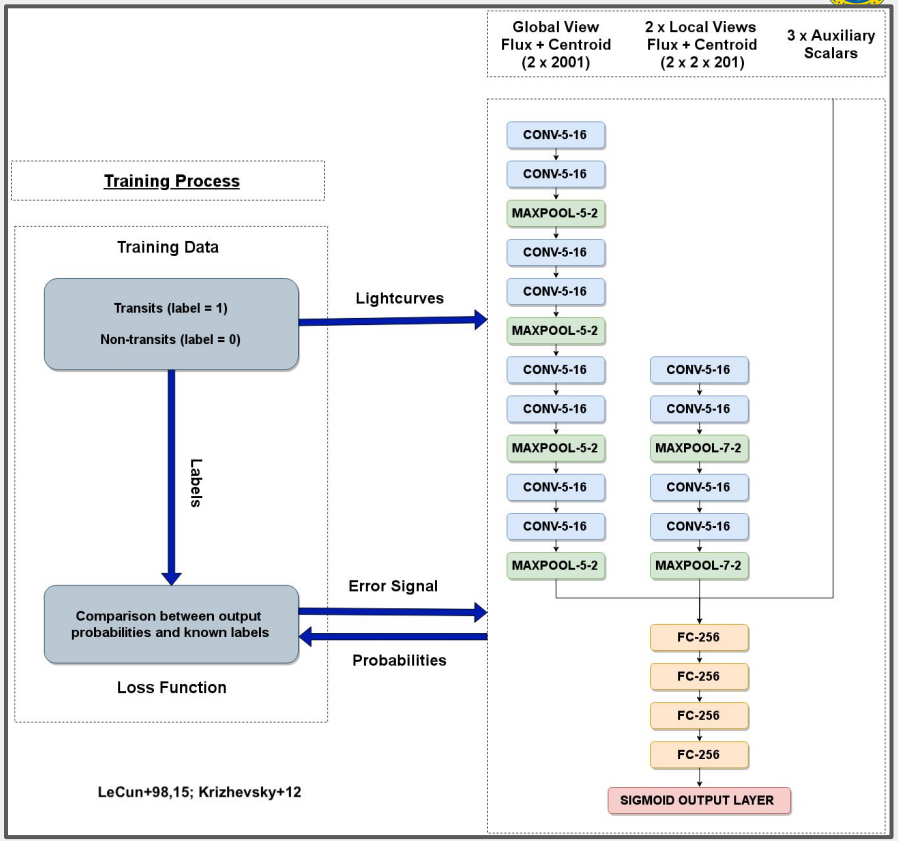
- Network based on the original neural network paper by Shallue et. al (2018)

- 890,000 lightcurves were searched from 91 fields

- 212,000 candidate transit detections using BLS from 58,500 stars

- Each candidate was manually vetted and assigned a label

- We had 15 planets

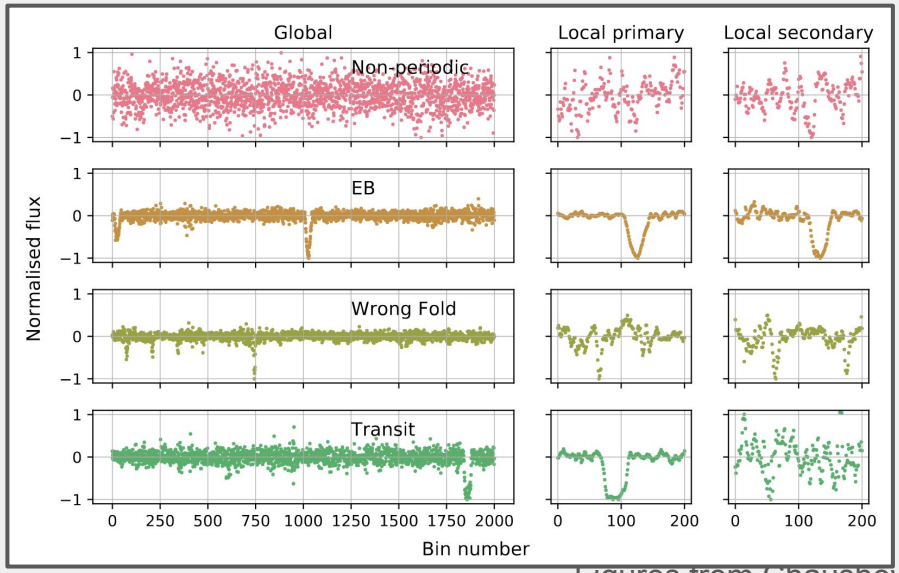


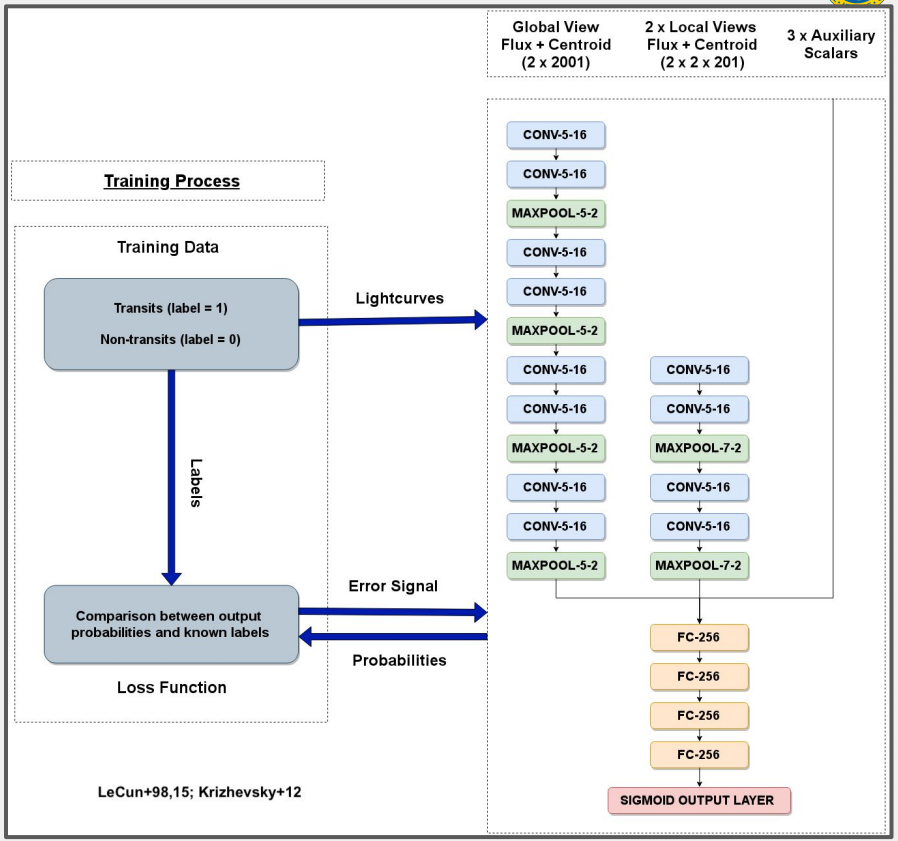
Figures from Chaushev et. al (2019)

Network Architecture + Data

- Network based on the original neural network paper by Shallue et. al (2018)

- Our dataset is heavily imbalanced: 14,000 FPs to 1 real planet! A balanced diverse dataset was critical for training





Figures from Chaushev et. al (2019)

Balanced FPs

Table 1 (Right): Training performance for different mixtures of false positives

Table 2 (Below): comparison to human vetters

Model	OFP selection	AUC	Accuracy	Precision	Recall
OFP	Max	0.992 ± 0.002	0.956 ± 0.006	0.960 ± 0.011	0.960 ± 0.011
	Min	**0.994 ± 0.000**	**0.964 ± 0.002**	**0.974 ± 0.003**	**0.974 ± 0.003**
	Uniform	0.993 ± 0.001	0.960 ± 0.002	0.968 ± 0.004	0.968 ± 0.004
	Random	0.993 ± 0.000	0.960 ± 0.002	0.967 ± 0.004	0.967 ± 0.004
NP/EB/OFP/WF	Max	0.958 ± 0.002	0.886 ± 0.002	0.902 ± 0.006	0.902 ± 0.006
	Min	0.954 ± 0.002	0.882 ± 0.002	0.906 ± 0.007	0.906 ± 0.007
	Uniform	0.955 ± 0.001	0.883 ± 0.002	0.905 ± 0.006	0.905 ± 0.006
	Random	0.958 ± 0.001	0.887 ± 0.002	0.907 ± 0.005	0.907 ± 0.005
NP/EB/OFP	Max	0.956 ± 0.002	0.885 ± 0.003	0.900 ± 0.006	0.900 ± 0.006
	Min	0.953 ± 0.001	0.881 ± 0.002	0.904 ± 0.006	0.904 ± 0.006
	Uniform	0.954 ± 0.002	0.882 ± 0.002	0.905 ± 0.006	0.905 ± 0.006
	Random	0.957 ± 0.001	0.886 ± 0.002	0.903 ± 0.005	0.903 ± 0.005
NP/EB		0.968 ± 0.001	0.903 ± 0.001	0.924 ± 0.004	0.924 ± 0.004
NP/EB/WF		0.958 ± 0.002	0.891 ± 0.002	0.908 ± 0.004	0.908 ± 0.004
NP		0.960 ± 0.001	0.901 ± 0.002	0.933 ± 0.005	0.933 ± 0.005

Model	AD	AS	BS	D	EA1	EA2	EB	OTH	SINE	No Flag
OFP	0.627 ± 0.027	0.321 ± 0.021	0.332 ± 0.026	0.302 ± 0.016	0.671 ± 0.028	0.796 ± 0.029	0.942 ± 0.010	0.870 ± 0.009	0.920 ± 0.007	0.877 ± 0.009
NP/EB/OFP	0.825 ± 0.011	0.485 ± 0.022	0.489 ± 0.023	0.515 ± 0.015	0.692 ± 0.014	0.848 ± 0.008	0.937 ± 0.003	0.771 ± 0.009	0.909 ± 0.006	0.767 ± 0.009
NP/EB/OFP/WF	0.855 ± 0.014	0.544 ± 0.025	0.521 ± 0.027	0.566 ± 0.018	0.677 ± 0.013	0.839 ± 0.008	0.949 ± 0.002	0.730 ± 0.011	0.926 ± 0.008	0.744 ± 0.011
NP/EB/WF	0.968 ± 0.003	0.726 ± 0.027	0.737 ± 0.026	0.805 ± 0.010	0.243 ± 0.009	0.298 ± 0.015	0.415 ± 0.007	0.229 ± 0.002	0.338 ± 0.008	0.413 ± 0.011
NP/EB	0.971 ± 0.002	0.774 ± 0.018	0.775 ± 0.020	0.836 ± 0.010	0.219 ± 0.009	0.282 ± 0.016	0.374 ± 0.008	0.214 ± 0.006	0.292 ± 0.008	0.378 ± 0.011
NP	0.996 ± 0.002	0.892 ± 0.015	0.861 ± 0.012	0.959 ± 0.005	0.006 ± 0.000	0.005 ± 0.000	0.021 ± 0.002	0.042 ± 0.005	0.100 ± 0.007	0.090 ± 0.005

CNNs are (relatively) insensitive to label noise

- We investigated network performance as a function a purposefully mislabeled data

- The network is able to ignore a small number of mislabelled candidates in the training process and still achieve good results!

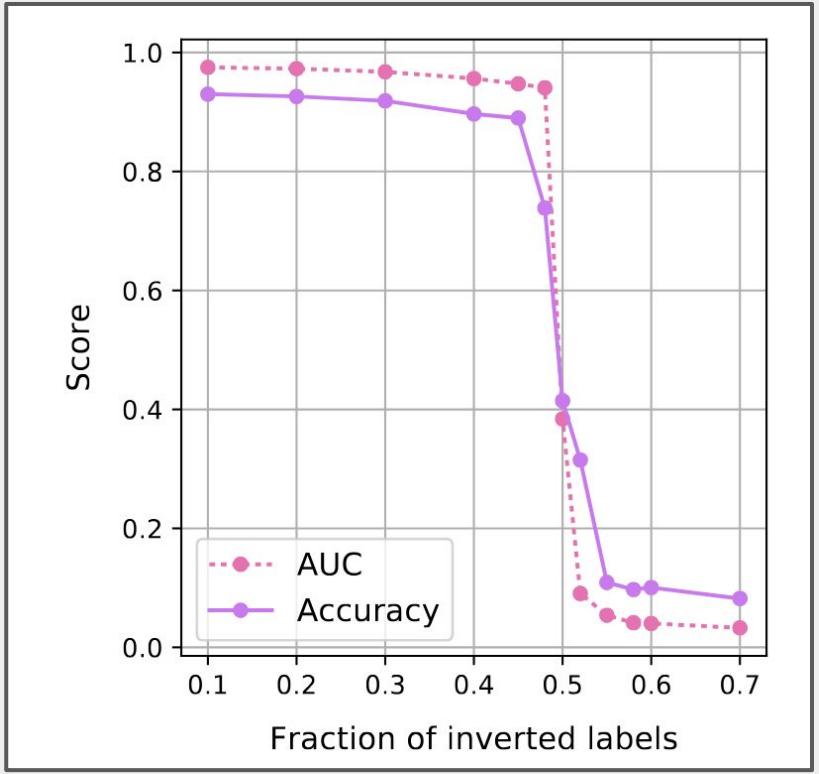


Figure from Chaushev et. al (2019)

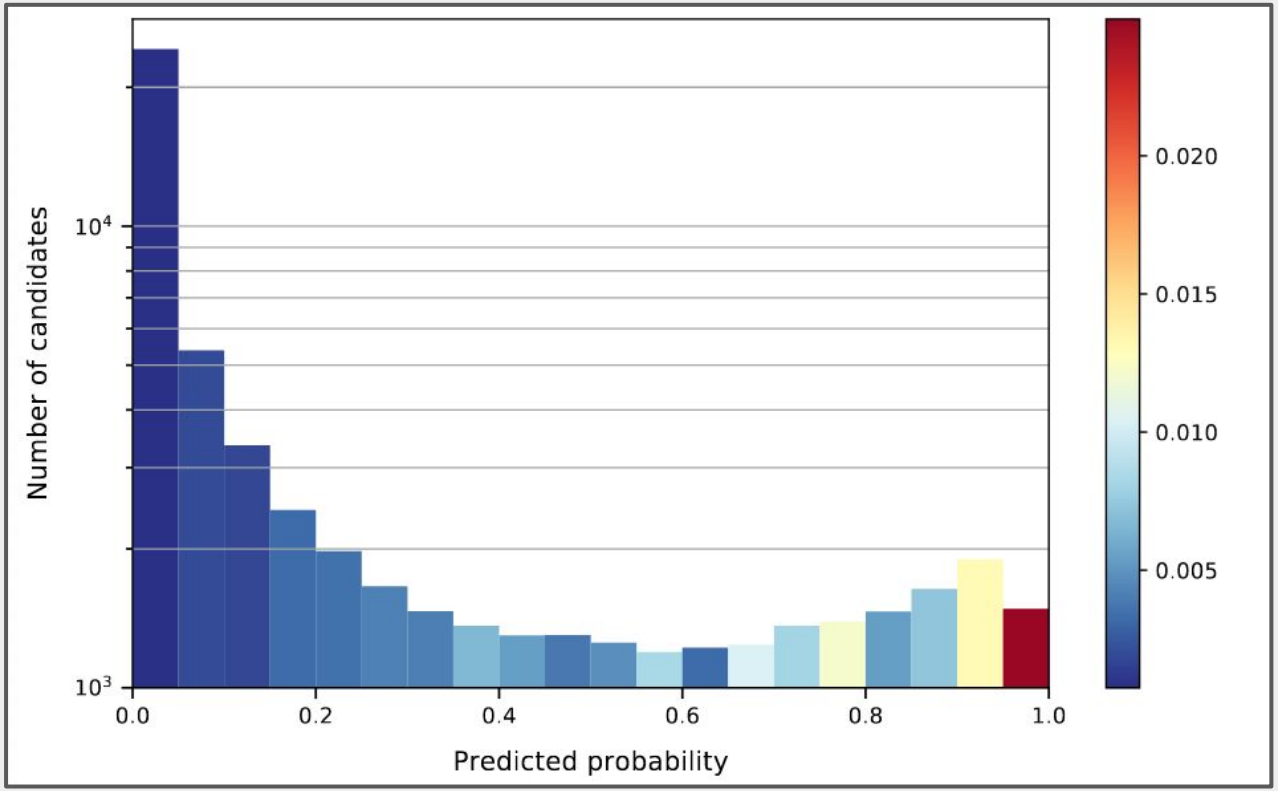
Results of applying classifier

- 890,000 lightcurves were searched from 91 fields

- 212,000 candidate transit detections using BLS from 58,500 stars

- Each candidate was manually vetted and assigned a label

- First iteration of neural network filters out 50% of false positives while recovering all good candidates and 14/15 planets



Figures from Chaushev et. al (2019)

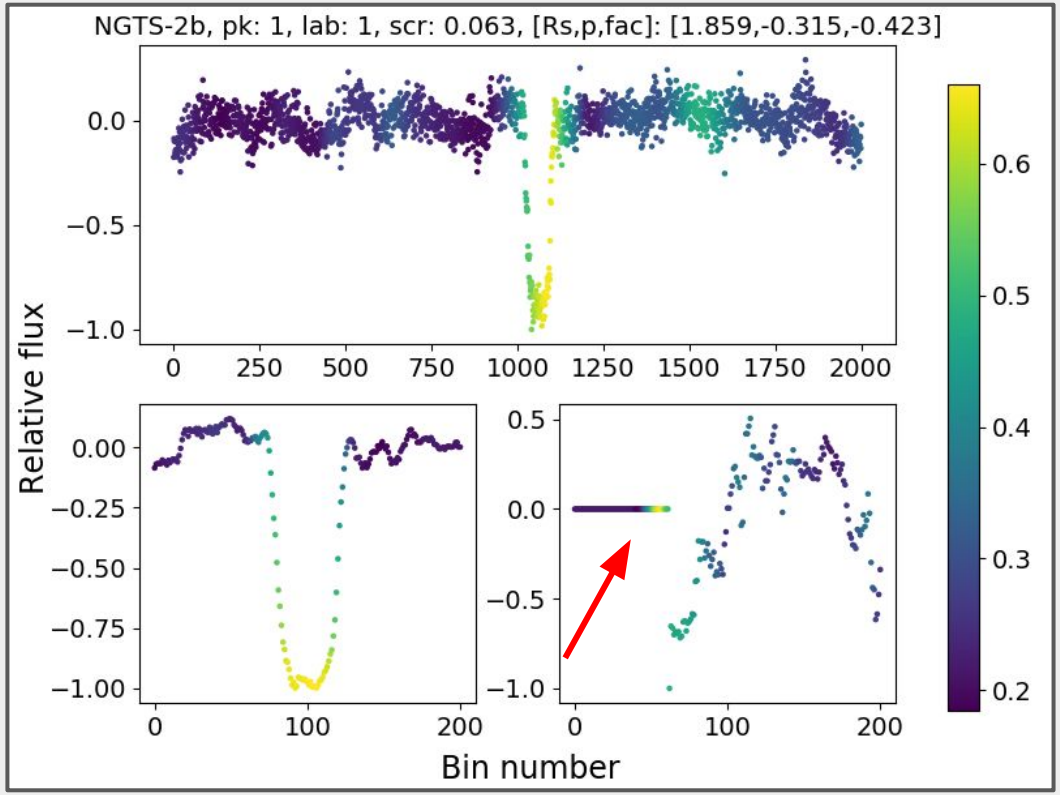
Why did we miss a planet?

- Feature activation map of the NGS-2b from GRAD-CAM++ (Chattopadhay+18)

- Network is activating on the discontinuity in the secondary transit present because of clipped data

- Replacing secondary with Gaussian noise yields a probability of 0.981

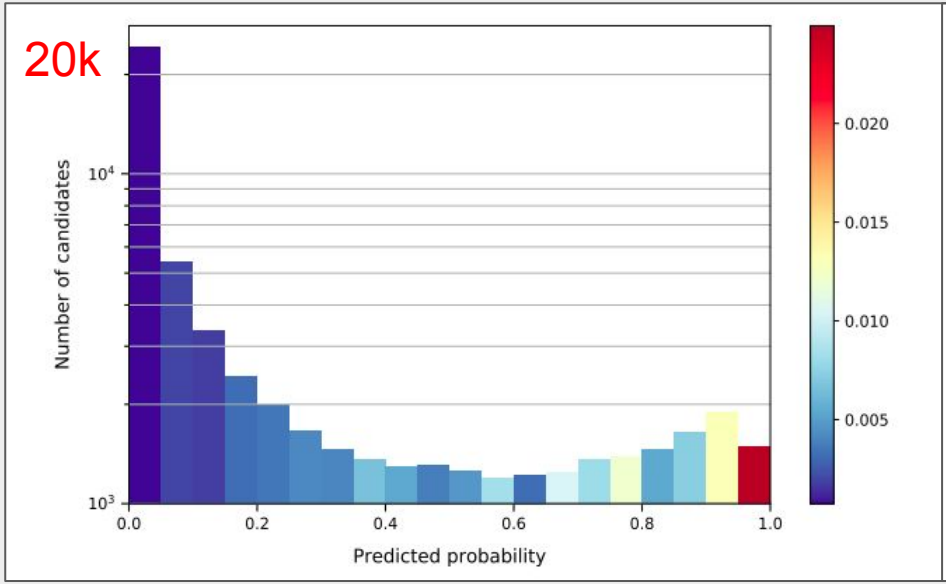
Solution: replacing missing data points with +1 instead of 0s



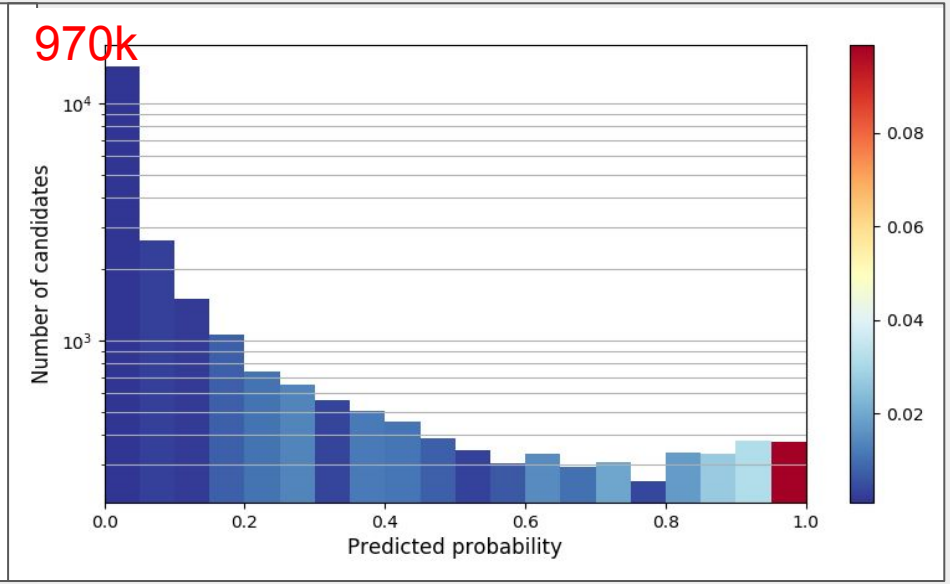
Figures from Chaushev et. al in prep

Performance with Dataset Size

Training Dataset Size	AUC	Accuracy	Precision	Recall
Paper I (20k)	95.8±0.2	88.6±0.2	90.2±0.6	90.2±0.6
Paper II (970k)	97.6	91.3	93.5	90.0



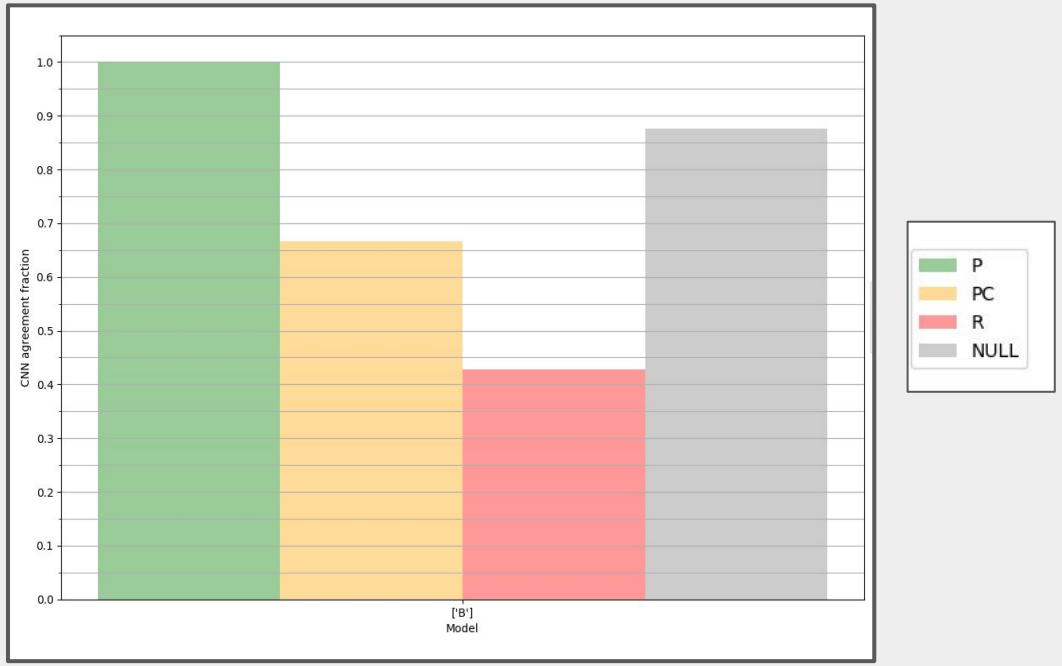
50% with probability < 0.1



80% with probability < 0.1

How does this compare to human vetters?

PLANET-ID	Network Score
103527	0.961
105452	0.997
101129	0.999
104605	0.997
101155	0.989
103351	1.000
101123	0.904
101101	0.956
103524	0.995
103867	0.998
105553	0.987
101631	0.995
101632	0.998
103406	0.998
103934	0.989



Comparing final assigned labels human + further follow-up (P: planet; PC: planet candidate, R: rejected, NULL: to be vetted) to CNN classifications

Backup Slides

A short history of extrasolar planets (exoplanets)

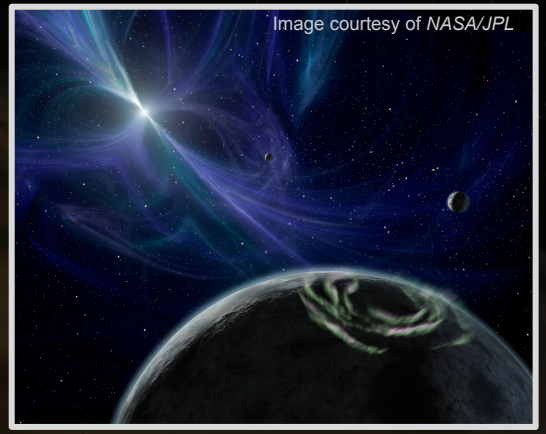
16th century - scientists speculate about the existence of exoplanets e.g.:

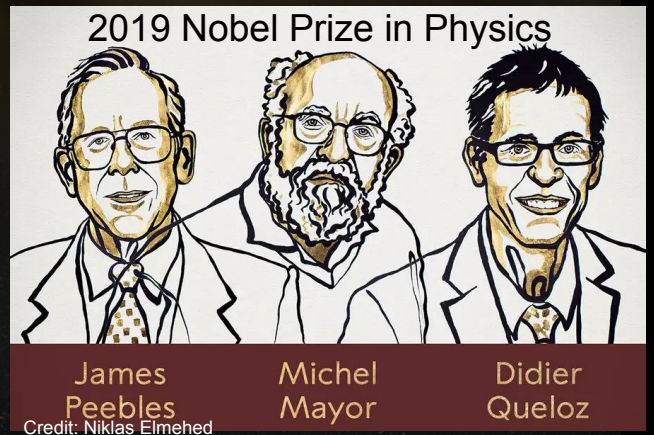
> *"...if the fixed stars are the centres of similar systems, they will all be constructed according to a similar design" - Newton*

1988 - First signs of an exoplanet spotted (but not confirmed until 2002!) around Gamma Cephei

1992 - Aleksander Wolszczan & Dale Frail discover first two exoplanets around pulsar (PSR) B1 257+12.

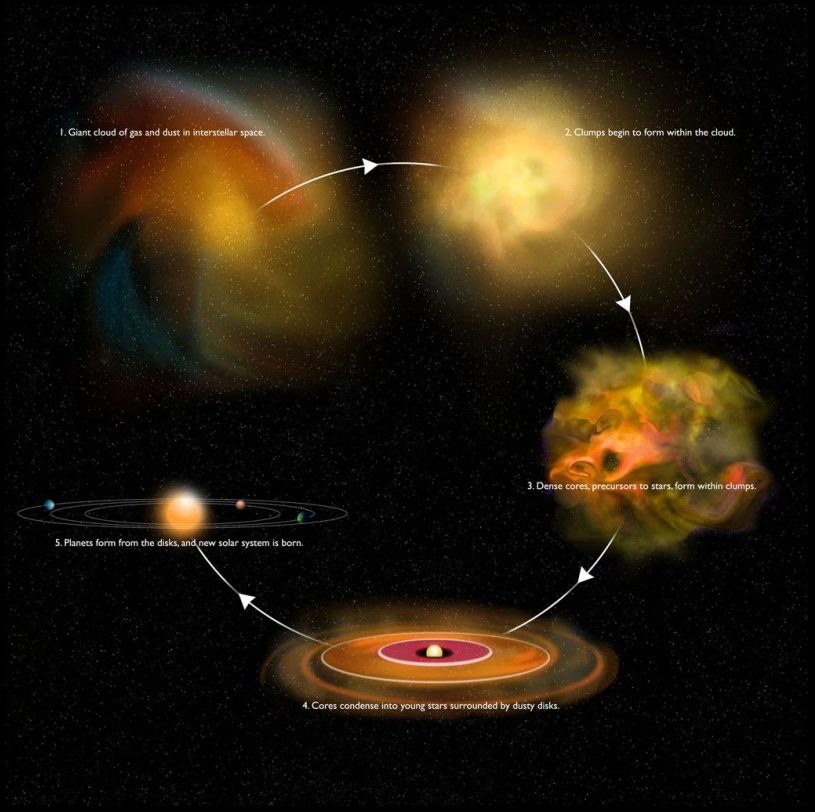
1995 - Michel Mayor and Didier Queloz find 51 Peg b, a hot jupiter in a 4.5 day period orbit around a main sequence star

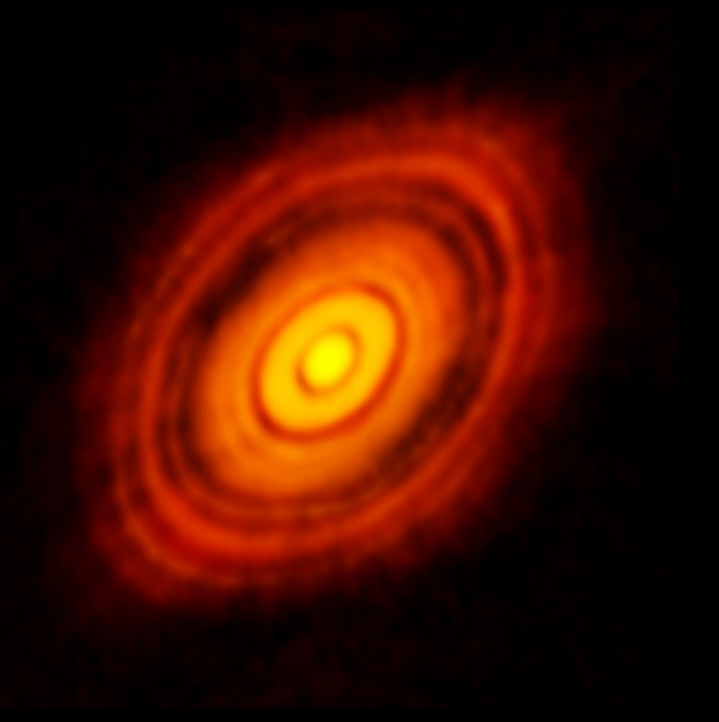




Credit: Niklas Elmehed

How do these planets form?





HL Tauri (150pc)
Observed with
ALMA