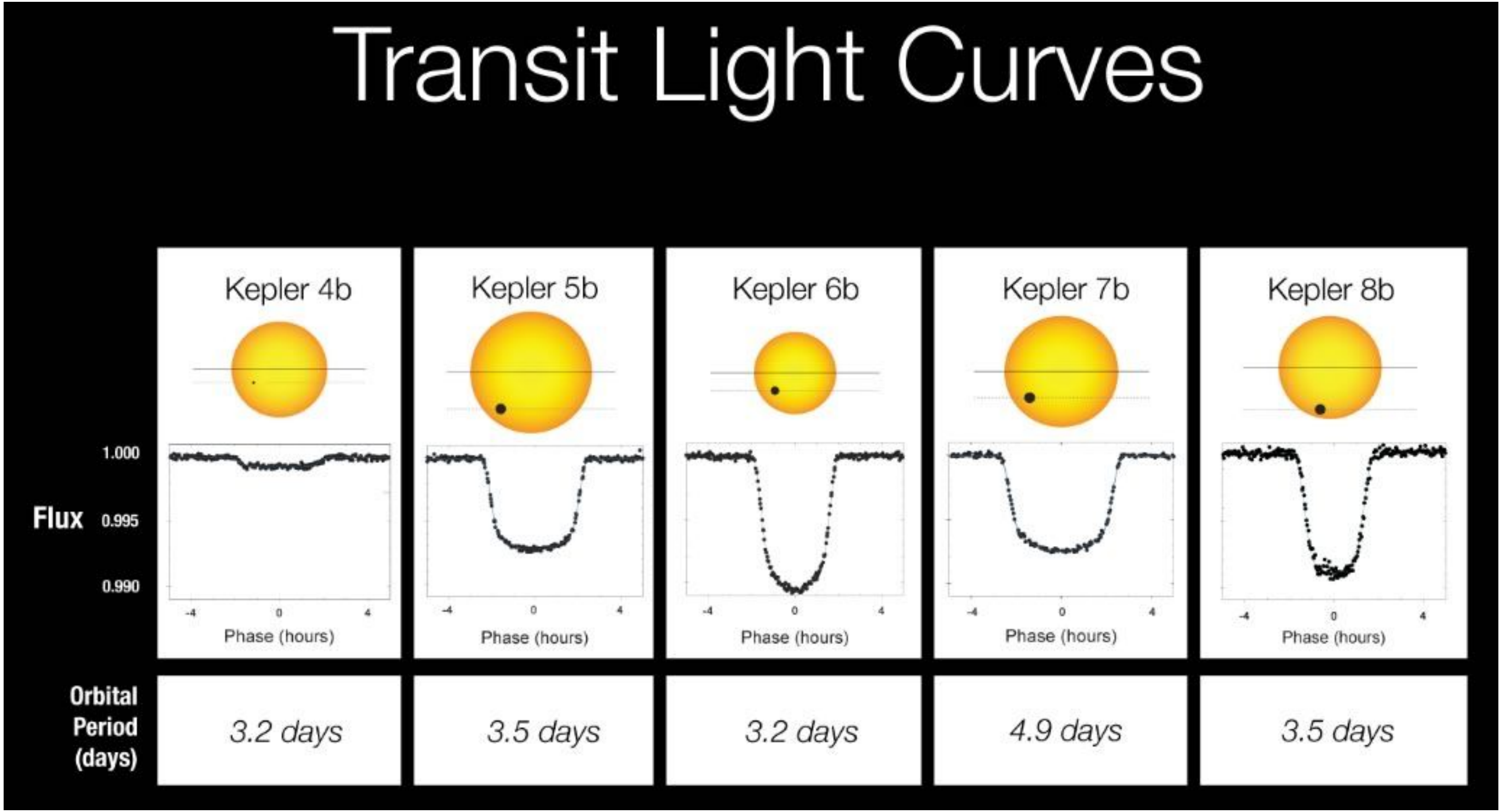
Exoplanet Detection using Machine Learning

Abhishek Malik
MSc Astrophysics - USM, LMU Munich
Supervisor: Dr. Ben Moster
a.malik@usm.lmu.de

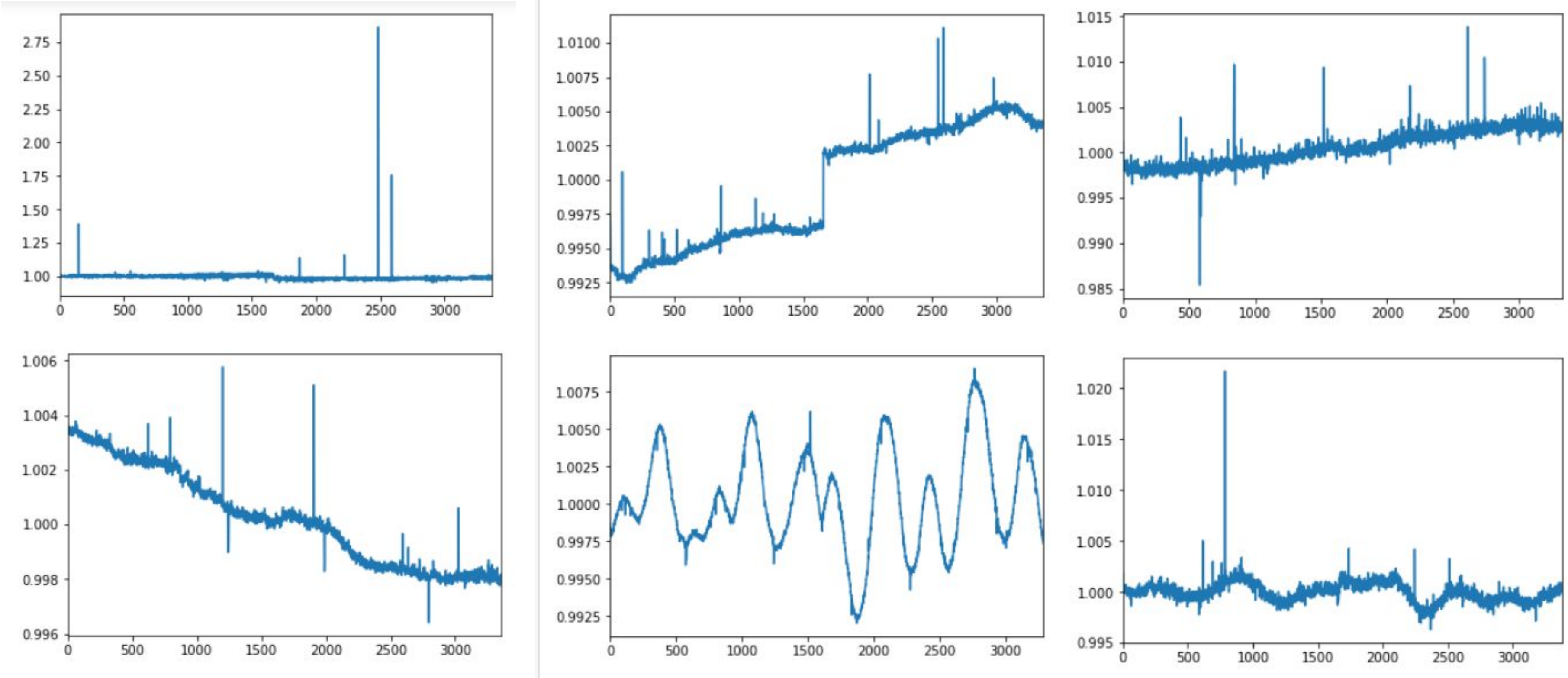
Contents:

- Introduction
- Data Preparation
- Results - simulated and real data
- Comparison with other similar models
- Exoplanet detection in 2020
- Conclusion
- References

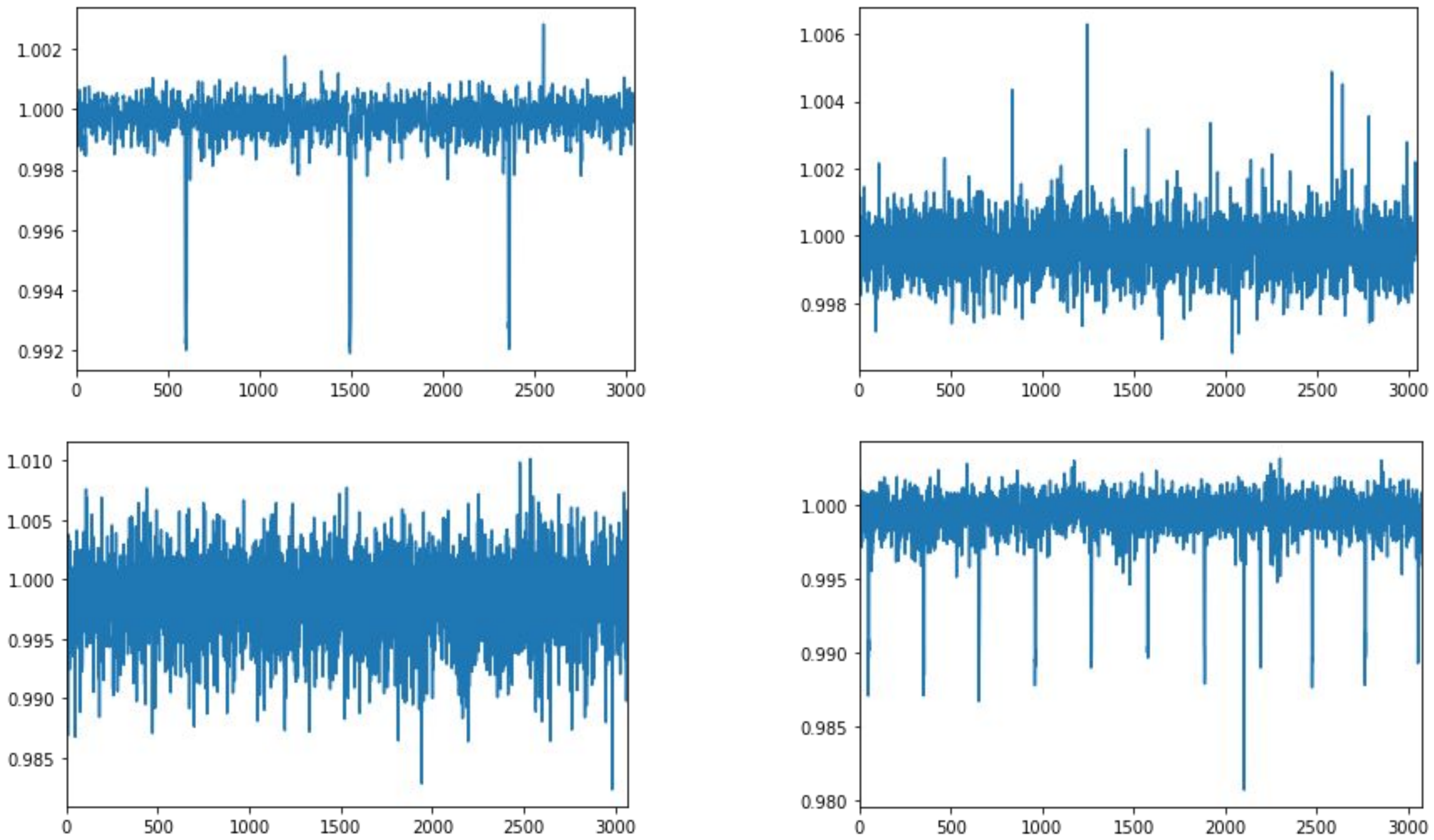
Identify transit dips in flux data (time-series) of the star



Raw lightcurves are messy:
Flux signals of few cases with planets



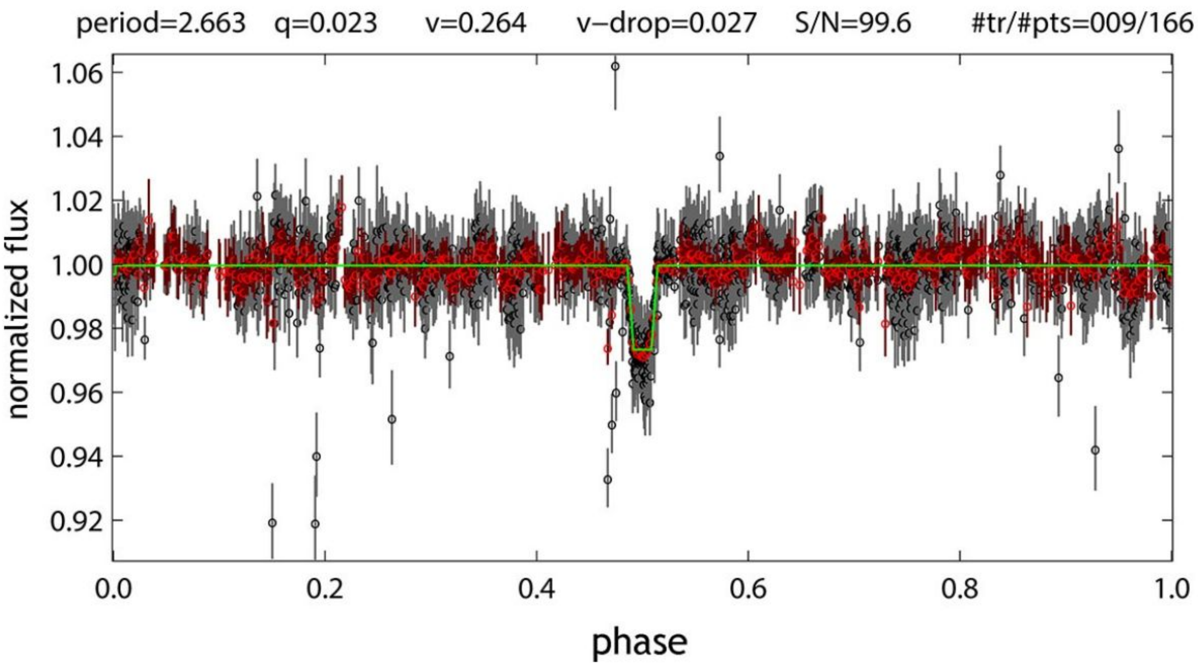
Cleaned data



Conventional method: BLS

BLS (Box fitting Least Squares):

- Usually works well
- Can fit to noise of the data => needs manual attention
- TESS produced over **1 million lightcurves each month**
- **Need a reliable system to automatically identify new candidates**

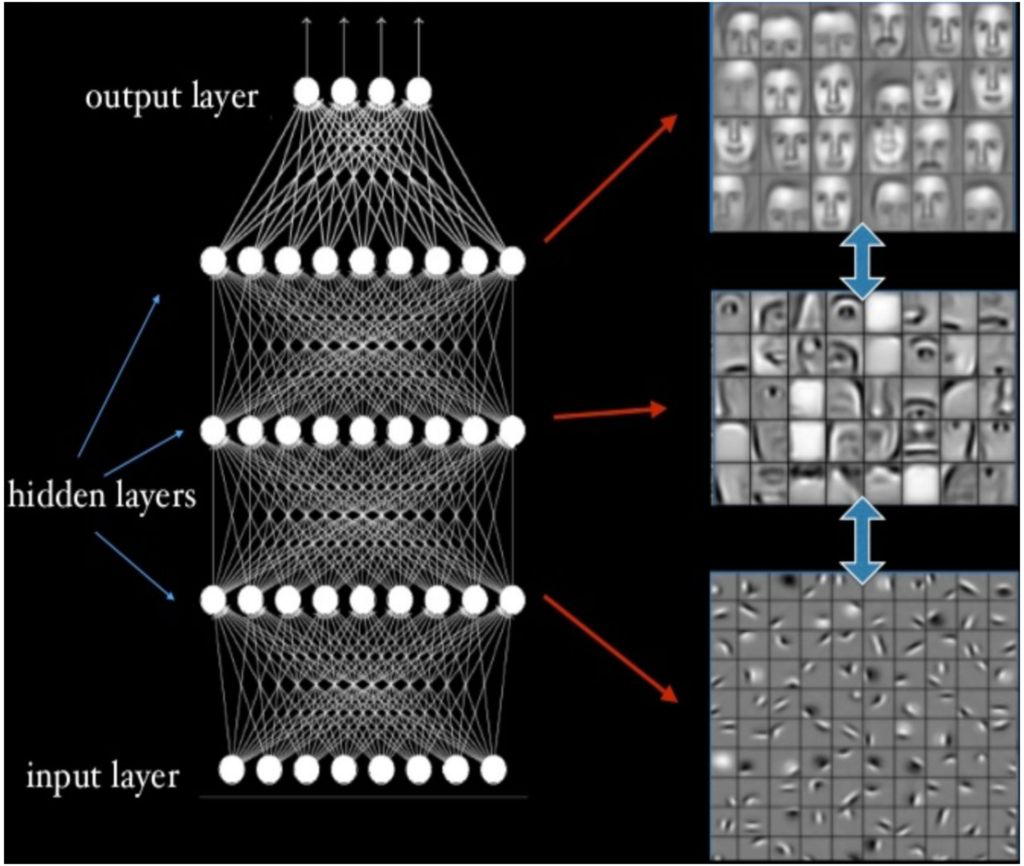


The Method:

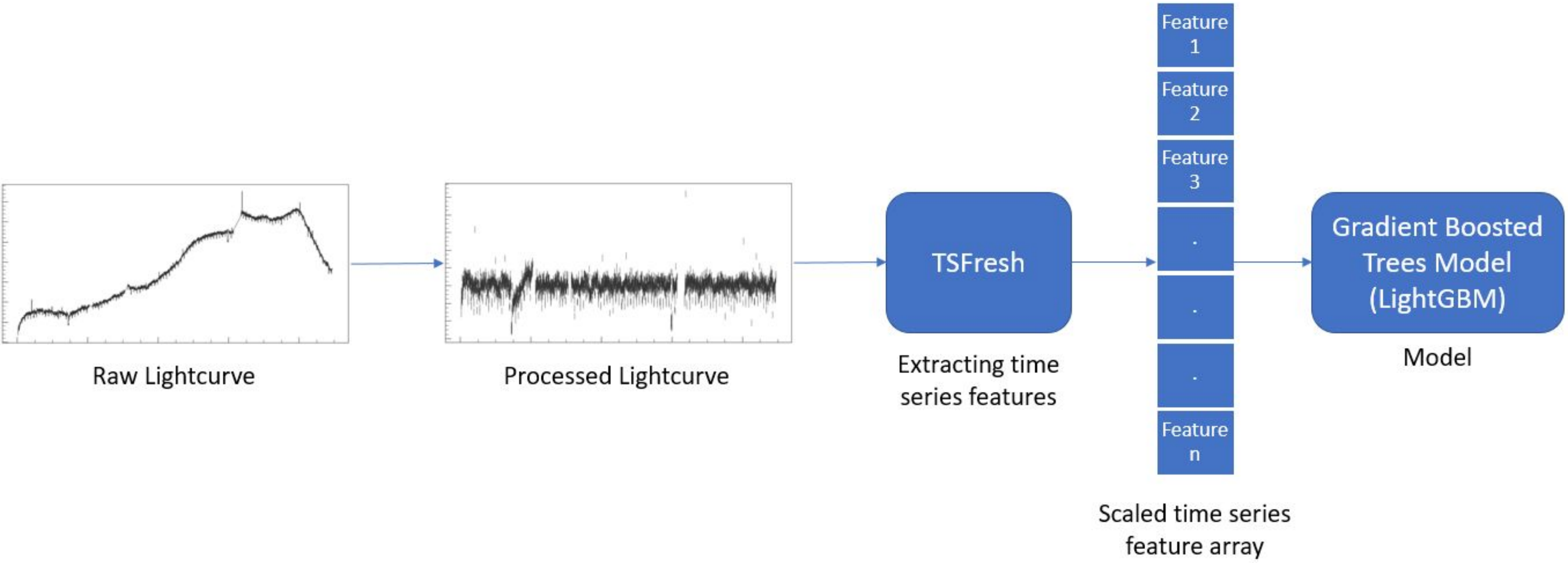
- Deep learning models can automatically generate features that are important for the task.

>> Extract general features and use them as input for a classical ML model.

>> Used a python library '*TSFresh*' to extract generalized time series features from each lightcurve.



Workflow:


Raw Lightcurve
Processed Lightcurve
TSFresh
Extracting time series features
Feature 1
Feature 2
Feature 3
.
.
.
Feature n
Scaled time series feature array
Gradient Boosted Trees Model (LightGBM)
Model

Simulated Dataset:

- Used data from the K2 mission.
- Removed all known planet signals from it
- Cleaned the lightcurves : Flattening, removing variability etc.
- Randomly injected transits (simulated planets) in 50% of the cases

>> Training set: ~6000 cases, Validation set: ~2000 cases

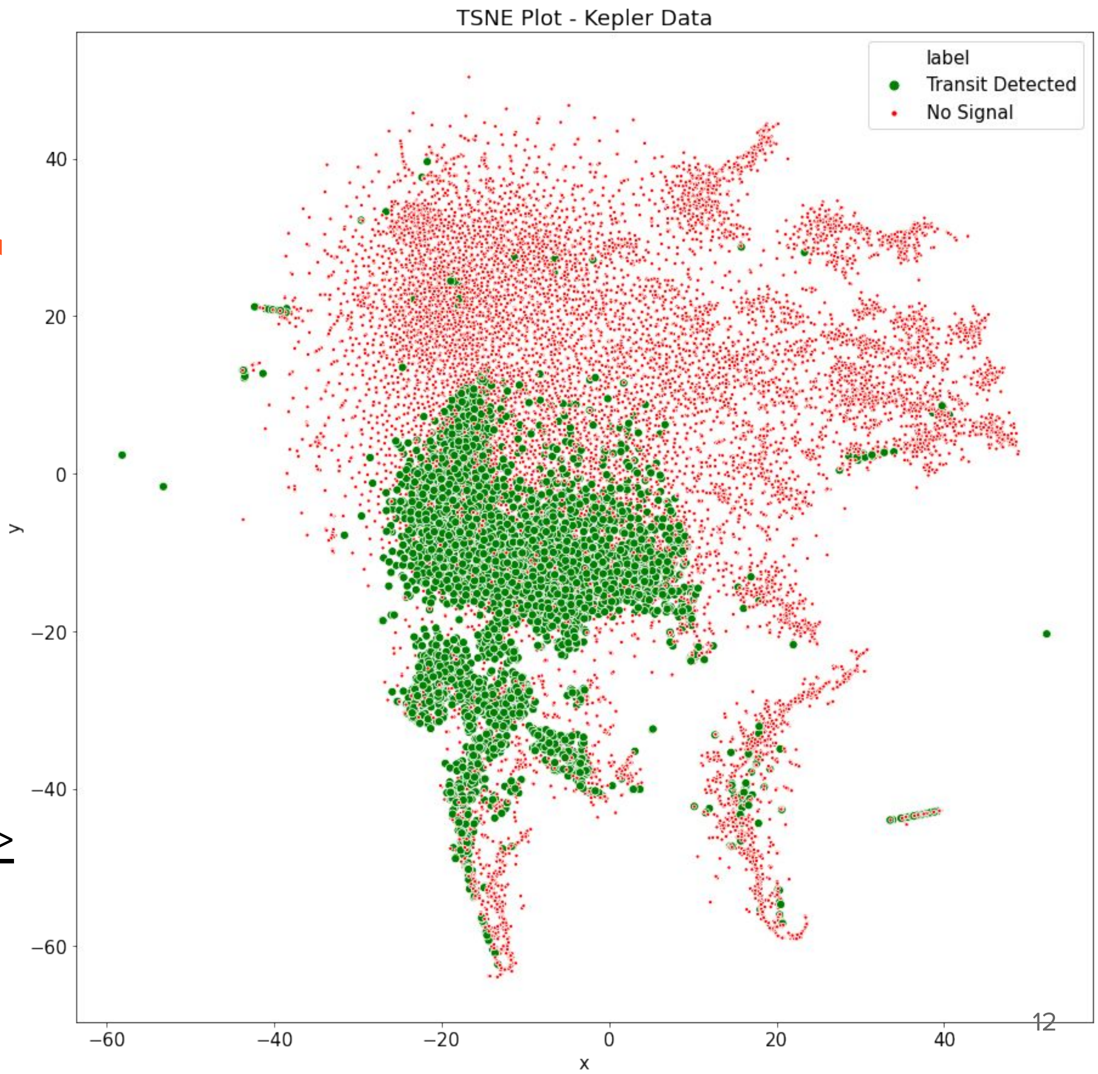
Results:

- The model was able to predict with an accuracy of **91%**.
- The model detected **92%** of all the planets with a precision of 94% (i.e. 6% false positives).
- Detected **84%** of all planets using BLS.
- This validated our proof of concept.

Application on Real Data: Kepler

- Real data from NASA's Kepler Archive
- Total 15737 lightcurves
- 3600 cases with confirmed planets

TSNE Plot >>



Results:

- **AUC: 0.948**. ~95% of the times it ranks a planet signals higher than false positive signals.
- **Recall: 0.96**: Model was able to identify 96% of all the planets.
- **Precision: 0.82**: Model produces 18% false positives.

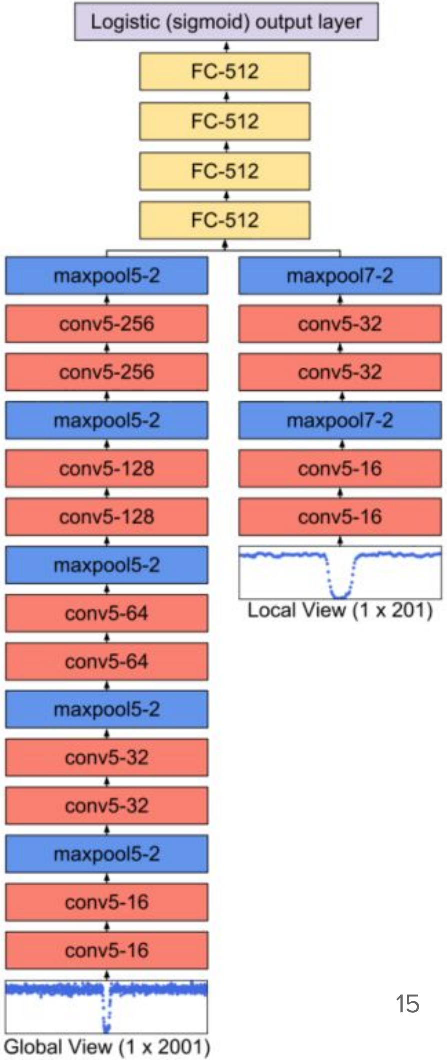
Comparison with other models - same data

Model	Dataset	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)
CNN	Processed	0.93529	0.96453	0.93839	0.93529
LSTM	Processed	0.93464	0.9652	0.93673	0.93464
Astronet (CNN)	Processed, local and global views	**0.960**	**0.988**	0.93	**0.95**
Astronet (MLP)	Processed, local and global views	**0.941**	**0.977**	N/A	**N/A**

Astronet (CNN)

Model Architecture >>

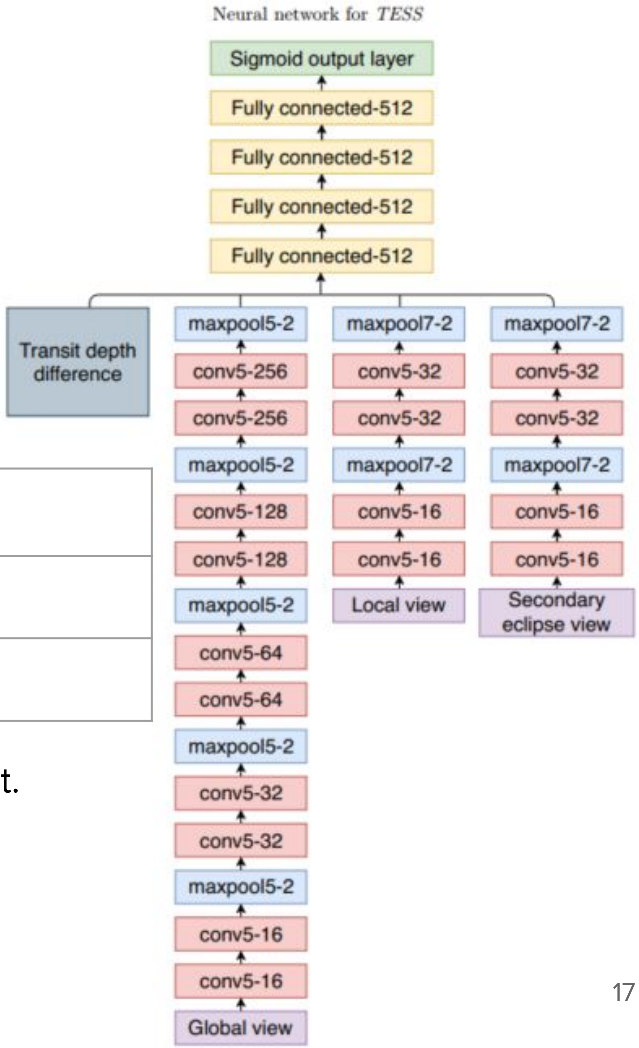
	AUC	Recall	Precision
Astronet	0.988	0.95	0.93
ML	0.948	0.96	0.82



Exoplanet detection in 2020: TESS

- TESS or Transiting Exoplanet Survey Satellite is the successor of kepler
- Launched in 2018 - Not many confirmed candidates
- Less TESS candidates => Highly unbalanced dataset:
  - 490 planets out of 16.5k cases

DL Model Architecture



	AUC	Recall	Precision
Astronet - TESS	NA	0.89	0.44
ML	0.80	0.82	0.81

Results of both models are likely to improve on a more balanced dataset.

Advantages of our methods:

- Less Processing: ML models can work only with global view; DL requires folded/secondary views.
- Faster: ML model **trains in < 5 minutes** on a normal PC; **DL models can take upto 5 hours** for training and much longer to hypertune. No GPUs required!
- Versatility: Exact same model setup / code can be used for different data sources eg. Kepler, TESS etc. DL models almost always needs to be changed with data source

Conclusion: Disadvantages

- Lower performance: DL models produce better results after proper training.
- Needs feature extraction: Time series data (eg. global view, folded view etc) can be directly used as an input in the DL model. While ML model requires extracted features from the time series data.

Conclusion:

- ML methods can lead to better results than conventional methods like BLS.
- Bigger is not always the better, smaller and simpler models can work very well when used properly.

Main References:

- Shallue C. J., Vanderburg A., 2018, The Astronomical Journal, 155, 94
- Catanzarite J. H., 2015, Technical report, Autovetter Planet Can-didate Catalog for Q1-Q17 Data Release 24. KSCI-19091-001 https://exoplanetarchive. ipac. caltech. edu/docs/KSCI-19091
- Dattilo A., et al., 2019, The Astronomical Journal, 157, 169 Friedman J. H., 2001, Annals of statistics, pp 1189–1232 Howell S. B., et al., 2014, PASP, 126, 398
- Yu L., et al., 2019, The Astronomical Journal, 158, 25 This paper has been typeset from a TEX/LATEX file prepared by the author.MNRAS000, 1–9 (2015)
- Obermeier C., et al., 2016, A&A, 587, A49
- Coughlin J. L., et al., 2016a, The Astrophysical Journal Supplement Series,224, 12
- Chaushev A., et al., 2019, Monthly Notices of the Royal Astronomical Society, 488, 5232

Thank You!!

Feedback / Questions?

Email: a.malik@usm.lmu.de