Towards Deriving Theories from Data: Frontiers for Model Inference in Astro-Geophysics

NASA AI Workshop, November 26, 2018

Victor Pankratius
Massachusetts Institute of Technology
Kavli Institute for Astrophysics and Space Research

Email: pankrat@mit.edu
Web: victorpankratius.com
Overview

• Discuss AI in science – now and in the future

• Based on two examples:
  • Astrophysics: Exoplanet search
  • Geophysics: Earth deformation, volcanoes
Exoplanet Search

Transiting Exoplanet Survey Satellite (TESS)

- Near all-sky survey
- Launched April 18, 2018
- Kepler mission follow-up, stars 10-100 brighter
- Expecting thousands of new exoplanets smaller than Neptune and potentially dozens that are comparable to our Earth
- Full frame images every 30 minutes, 200,000 pre-selected stars monitored with 2 min cadence
- TESS processing pipeline extracts light curves
- Problems similar to future Big Data applications, e.g., Large Synoptic Survey Telescope (LSST) and others

[https://tess.mit.edu; https://tess.gsfc.nasa.gov; Ricker14]
Exoplanet Search

Transit Search: State-of-the-art

Unfolded Time Series

Folded Versions for Transit Search

Parameters

- KIC 9458613 (Kepler 33)
- Normalized Flux vs. Time
- Period: 13.1756...
- 31.7844...
- 41.029...
- 21.775...
- 5.6679...

→ Machine learning and other methods typically applied on folded light curves [Shallue18]

[Kovacs02, Seager11, Winn14]
Exoplanet Search

However, there is more information in the unfolded time series.

→ Revealing irregular Transit Timing Variations (TTV) in Kepler90 system

“Zooming” in on transits; red & black lines = catalog-listed periods

“Year” of Kepler90g is 1 day longer in this particular transit!
Bi-directional LSTM Networks in Exoplanet Search

A Toy Example:

Networks that are “deep” in time
Bi-directional LSTM Networks in Exoplanet Search

BDLSTM example: learning **planet transits**

[training: 50 epochs, 1 second steps, 0.5 dropout rate, until accuracy = 0.9797]

Applying trained BDLSTM to other light curves

[Images showing light curves for KeplerID 2581316, 11442793, and 3247268]
Bi-directional LSTM Networks: Other Phenomena

Variable Star Phenomena: Learning Dwarf Nova Events

Example: V344 Lyr (Kepler 7659570)

- Training set = 1 piece of time series
- Preliminary BDLSTM Prediction on Test Set (rest of time series)

Note: potentially useful prediction capability based on empirically learned model
Next: Establishing Data – Model Connections

What do humans typically do?

- Look at light curve → develop a “mental model”
  (hypothesized planetary system, related phenomenon)
- “Play” in imagination, unfold over time
- Anticipate dynamics
- Look back at the light curve for supportive clues

→ Inverse problem solved iteratively by generating multiple forward models + pruning those that do not exhibit the right properties

→ This process can be automated
Next: Establishing Data – Model Connections

Proof of concept example:

Programmatic Interface in Python Jupyter Notebooks

blender.org Raytracer
Generative Approach

Generate Physical Model

- Scenario: One planet
- Scenario: Two planets
- Scenario: Irregular Orbiting Debris
Generative Approach: One Planet

A “Rosetta Stone” linking models & theories to data

Theoretical Domain

Physics

Model Features

Data Features

Empirical Domain

+ noise
Generative Approach: Two Planets

Theoretical Domain

Physics

Model Features

Data Features

Empirical Domain

+ noise

Relative Flux

Time

+ noise

Time
Generative Approach: Irregular Debris

Theoretical Domain

Physics → Model Features → Data Features

Empirical Domain

+ noise
Adding Inference Capabilities

A system with a confirmed planet might have other planets, moons, debris disks, …

→ create an “autocomplete” capability (inference engine) for planetary systems

Model from empirical data
Adding Inference Capabilities

A system with a confirmed planet might have other planets, moons, debris disks, …

→ create an “autocomplete” capability (inference engine) for planetary systems

→ “Guess where & what” with plausible physics

Model from empirical data
Adding Inference Capabilities

A system with a confirmed planet might have other planets, moons, debris disks, …

→ create an “autocomplete” capability (inference engine) for planetary systems

→ “Guess where & what” with plausible physics

→ Create a population of forward models and plausible variants (e.g., using genetic programming)

Derive empirical features to look for, if models were describing reality

• Generate neural networks that have higher attention in those areas
• Test / falsify multiple theories in parallel
Adding Inference Capabilities

Generative approach facilitates inference on other properties

Planet mass, radius, orbital parameters, rotation rate, obliquity
⇒ gravitational acceleration
⇒ atmosphere parameters
⇒ potential mean density/rockiness
⇒ inferences on core, magnetosphere.

Planet surface temperature
⇒ greenhouse warming
⇒ thermal emission
⇒ atmospheric gases and compositions.

Spectroscopy parameters
⇒ biosignatures, gases
⇒ indicator factors of habitability

Host star properties
⇒ luminosity/temperature, spectral type, activity, rotation rate, and flare activity
⇒ habitability
Can this approach can be transferred to other domains?
Geophysics Example

Volcanology

GPS Sensors → Time Series → Empirical Model

Empirical Model → Theoretical Model → Classifier for Earth deformation/ inflation event

Classified for Mogi Source


[Hibert et al., GRL ‘15]
Inferring Models at Higher Abstraction Levels

AI Theorem Prover for Science Models / Test Case Generator for Empirically Observable Features

- Derive test cases: “this property should be observable if this model was right”
- Derive falsification cases: “property that should never be observed if this model was right”
- Derive invariants: “this predicate should always be true if this model was right”
Symbolic Model Manipulation: Algebraic Approach

\[ M_{seed} = M_1 \oplus M_2 \oplus M_3 \oplus M_4 \oplus M_5 \]

\[ \Psi(M_{seed}) = \Psi(M_1 \oplus M_2 \oplus M_3 \oplus M_4 \oplus M_5) \]
\[ = \Psi(M_1) \oplus \Psi(M_2) \oplus \Psi(M_3) \oplus \Psi(M_4) \oplus \Psi(M_5) \]

\[ \mathcal{E}(M_{seed}, M_6) = \mathcal{E}(M_1 \oplus M_2 \oplus M_3 \oplus M_4 \oplus M_5 \oplus M_6) \]
\[ = M_1 \oplus M_2 \oplus M_3 \oplus M_4 \oplus M_5 \oplus M_6 \]

\[ \mathcal{I}(M_{seed}) = \mathcal{I}(M_1 \oplus M_2 \oplus M_3 \oplus M_4 \oplus M_5) \]
\[ = M_1 \oplus M_2 \oplus M_3 \oplus M_4 \]

\[ \mathcal{G}(\text{space}(M)) = M_i \text{ with } M_i \in \text{space}(M) \]

Remark: more elaborate modeling requires introduction of a type system, constraints / domain-specific rules, …

[Pankratius et al., AGU'18]
Examples for $M_i$ in Geoscience

Compute Interferogram

Test with Reality

Compare with real-world InSAR satellite or UAV interferogram

add machine-learned noise components

vertical deformation from expansion of two Mogi sources

Genetic Programming in Python, with a scikit-learn inspired API:

Test with Reality

[Rude, Pankratius, Rongier: work in progress]
• Where do we go from here?
Blueprint for “Astra”
An AI Science Assistant with Domain Knowledge
Conclusion

• Big Data & instrument fusion in scientific applications
  → push for more automation at all levels

• We need to rethink automation in the scientific process

• Problems go beyond detection, classifications, statistics

• Automated insight generation will be key

• Vision for future:
  AI science assistants that have domain knowledge
Thanks!

@vpankratius
pankrat@mit.edu
victorpankratius.com

Acknowledgement: This material is based upon work supported by the NSF and NASA. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the NSF or NASA.