Computer-Aided Discovery: Leveraging Machine Intelligence for Scientific Insight Generation

Victor Pankratius
Massachusetts Institute of Technology
Haystack Observatory
Head of Astro- & Geo-Informatics Group

@vpankratius
pankrat@mit.edu
Data – all the time – everywhere on earth and in space

- **Scalable machine assistance** is needed to help humans in the discovery process
- **Overcome human cognitive limits and bias** through algorithmic support
- The scientific discovery process becomes a search process across multidimensional data sets.
- **New Field:** Scientific question answering

---

Software-based Instruments / Backends

- Algorithms
- Parallel Computing
- Search, Classification
- Signal Processing
- Imaging
- Simulations
- Software Engineering
- Data Mining
Computer-Aided Discovery

Overcome cognitive limits with machine assistance

“Syntax”
Data

“Semantics”
Data Exploration, Analysis, Mining

“Pragmatics”
Discovery and Insight Generation

• What does feature X imply?
• How does it fit into the theoretical context?
• Does it contradict or confirm established models?

“This is feature X”, “class Y”, ...

Numbers correctly encode actual measurements

Semiotic Layers

Big Data / Real-World Phenomena

Victor Pankratius – Astroinformatics 2016
Computer-Aided Discovery: Overview

Instrument 1

Physics

Plasmasphere: Satellite Imager Data

Empirical Data

Data / Compute Center

Instrument 2

Cloud-Based Analysis Interface

Scientist

Search results – would you like to find more like:

Include scientist feedback

Model 1

Model 2

Model 3

Model 4

Artificial Intelligence

Theory
Example: Choices in Processing Workflows

- Choice of workflow processing parameters can affect the visibility and discovery of interesting phenomena.

Figures showing geophysical phenomena (global TEC views) over the northern hemisphere on 2014-02-27
Example: Choices in Processing Workflows

- Spectral energy distribution (SED) of galaxies
- Eigenspectra determined from a sample of 2000 SDSS galaxies selected at random; different choice and parameters of Principal Component Analysis to describe variability of data

Information and figure obtained from Walcher J., et al. 2011, Astrophys Space Sci, 331, 1
Example: Choices in Processing Workflows

Astronomical Object Detection

<table>
<thead>
<tr>
<th>Article</th>
<th>Image transformation</th>
<th>Image type</th>
<th>Aim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herzog &amp; Illingworth</td>
<td>Mean</td>
<td>Optical</td>
<td>SD</td>
</tr>
<tr>
<td>Le Fèvre et al.</td>
<td>Bijouei</td>
<td>Multi-band</td>
<td>SD</td>
</tr>
<tr>
<td>Stetson</td>
<td>$\sigma$-clipping + Gaussian</td>
<td>N/A</td>
<td>SD</td>
</tr>
<tr>
<td>Slezak, Bijouei &amp; Mars</td>
<td>Gaussian + Bijouei</td>
<td>Optical</td>
<td>SD</td>
</tr>
<tr>
<td>Bertin &amp; Arnouts</td>
<td>$\sigma$-clipping</td>
<td>N/A</td>
<td>SD</td>
</tr>
<tr>
<td>Szalay, Connolly &amp; Szokoly</td>
<td>Gaussian</td>
<td>Multi-band</td>
<td>FSD</td>
</tr>
<tr>
<td>Mighell</td>
<td>Mean</td>
<td>N/A</td>
<td>SD</td>
</tr>
<tr>
<td>Hopkins et al. 2002</td>
<td>Gaussian</td>
<td>Radio</td>
<td>SD</td>
</tr>
<tr>
<td>Aptona, Lefèvre &amp; Collet</td>
<td>Morphological</td>
<td>Multi-band</td>
<td>SD</td>
</tr>
<tr>
<td>Yang, Li &amp; Zhang 2008</td>
<td>Median</td>
<td>Optical</td>
<td>SD</td>
</tr>
<tr>
<td>Perret, Lefèvre &amp; Collet</td>
<td>$\sigma$-clipping + Median</td>
<td>Multi-band</td>
<td>SD</td>
</tr>
<tr>
<td>Haupt, Castro, &amp; Nowak</td>
<td>Distilled sensing</td>
<td>Radio</td>
<td>SD</td>
</tr>
<tr>
<td>Lang et al. 2010</td>
<td>Median</td>
<td>Multi-band</td>
<td>PSD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Article</th>
<th>Strategy</th>
<th>Image type</th>
<th>Aim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jarvis &amp; Tyson 1981</td>
<td>Local</td>
<td>Optical</td>
<td>FSD</td>
</tr>
<tr>
<td>Irwin 1985</td>
<td>Optical</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Le Fèvre et al. 1996</td>
<td>Local</td>
<td>Multi-band</td>
<td>SD</td>
</tr>
<tr>
<td>Slezak et al. 1988</td>
<td>Gaussian</td>
<td>Optical</td>
<td>SD</td>
</tr>
<tr>
<td>Bijouei &amp; Rui 1995</td>
<td>Gaussian</td>
<td>Optical</td>
<td>SD</td>
</tr>
<tr>
<td>Bertin &amp; Arnouts 1996</td>
<td>N/A</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Szalay et al. 1999</td>
<td>Global</td>
<td>Multi-band</td>
<td>PSD</td>
</tr>
<tr>
<td>Stark et al. 1999</td>
<td>Global</td>
<td>Mid-infrared</td>
<td>PSD</td>
</tr>
<tr>
<td>Lazzati et al. 1999</td>
<td>Global</td>
<td>X-ray</td>
<td>SD</td>
</tr>
<tr>
<td>Hopkins et al. 2002</td>
<td>Global</td>
<td>X-ray</td>
<td>SD</td>
</tr>
<tr>
<td>Freeman et al. 2002</td>
<td>Global</td>
<td>X-ray</td>
<td>SD</td>
</tr>
<tr>
<td>Makovoz &amp; Marleau 2006</td>
<td>Global</td>
<td>Multi-band</td>
<td>PSD</td>
</tr>
<tr>
<td>Melin et al. 2006</td>
<td>Local</td>
<td>Radio</td>
<td>SD</td>
</tr>
<tr>
<td>Yang et al. 2008</td>
<td>Optical</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Herranz et al. 2009</td>
<td>Global</td>
<td>Radio</td>
<td>PSD</td>
</tr>
<tr>
<td>Stark et al. 2009</td>
<td>Gamma-ray</td>
<td>SD</td>
<td></td>
</tr>
</tbody>
</table>


SD: source detection
FSD: faint source detection
PSD: point source detection
ESD: extended source detection
Computer-Aided Discovery: Conceptual Approach

Problem Space
Domain rules
Constraints
Settings
Inputs/outputs

Theory Representation Space
Synoptic Model Representation

Basic Model
Sub-Model 1
Sub-Model 2

<<contains>>

<<optional >>

<<alternative >>

Parameters a: Range [1..10], b=0
Preprocessing: Algorithm_1

Parameters a: Range [30..40], b=5
Preprocessing: Algorithm_2

Feedback/Refinement/Insight

Empirical Evaluation Space

Data Set | 1 | 2 | ..
---------|---|---|---
Theoretical Model Variant 1
Theoretical Model Variant 2
Theoretical Model Variant n
Computer-Aided Discovery: Infrastructure

“Sensors”
- Generate compute jobs
- Move computation to data locations

Data & Compute Cloud

Service Layer
- Computer-Aided Discovery Engine
  - Execution Crawlers
  - Web Service APIs
  - Direct service integration into other tools

Cloud-Based Analysis Interface

Scientist

Data Integrators, indexing, etc.

Scikit Data Access (Python)
https://github.com/skdaccess/skdaccess
(Kepler, USGS, and other data sources)
Application Examples

Ionosphere & Space Weather

Geospace TEC – Imaging Workflow

Raw GPS L1/L2 data → Input Filtering → Spatial Filtering → Structure Recognition

- Station Selection
- Bias Estimation
- Background Estimation
- Binning
- 2D Fourier Filter
- Gradient Detector
- Minimum Scalloping
- Local Inverse Filter
- Model-based
- Functional Smoothing
- Windowing Function 1
- Windowing Function 2

Variants of ionospheric feature models

Application Examples

Geoscience / Volcanics

Variants of earth deformation models

Variants of magma source models
(5 different source model inversions)

Component Analysis

[54.22°N, 54.19°N, 54.1°N, 54.04°N]
[166°W, 165.85°W, 165.7°W]

[6.1°W, 6°W, 5.9°W, 5.8°W, 5.7°W]

[54°N, 54.1°N, 54.2°N, 54.3°N]

AV06

AV07

AV08

AV10

AV12

AV14

AV15

54°N

54.1°N

54.2°N

54.3°N

0.0

0.4

Similarity Metric:
Summed Mogi
Vector Distance [mm]

Model Displacement
PCA Displacement

20%

Sill

Constant

Open Pipe

Closed Pipe

Mogi

Finite Sphere

Application Examples

Groundwater Studies

GPS (North, East, vertical)

+ Data Fusion: GRACE, ...

- Group wells by behavior
- Red groups represent wells whose water levels are highly correlated in time

[C.Rude, J.Li, M.Gowanlock, T.Herring, V.Pankratius, AGU 2016]
Application Examples

Planetary Science / Site Selection

Variants of combined site selection models

Plotted Over Overall Rank (darker = better)

[D.Blair, M.Gowanlock, J.Li, C.Rude, T.Herring, V.Pankratius, LPSC 2016]
Application Examples

Solar Science

Wavelet Based Characterization of Low Radio Frequency Solar Emissions

The left panel shows an example pre-processed dynamic spectrum from the MWA which has a resolution of 0.5 s and 40 kHz. The circles mark the features identified by our wavelet decomposition technique. The right panel shows the distribution of the detected features in the $\Delta t - \Delta \nu$ plane on a log scale.

Application Examples

Astronomy / Exoplanets

Kepler Space Telescope

Variants of hypothesized planetary systems

Figure from Shporer A., et al. 2011, AJ, 142, 195
Application Examples

Astronomy / Exoplanets

Light Curve Model Generation Example
**Application Examples**

**Astronomy / Lightcurves**

- **Multiple star systems:**
  - triple, ternary, quadruple, quintuple, sextuple, septuple ...
- **Exo-Moons**
- **Exo-Trojans** (small bodies in planetary Lagrangian points)
- **Variable stars**
- ...
Application Examples

Astronomy

...ongoing work

• Light curves
• Pulsars
• Variable stars
• Stellar flares
• Stellar population synthesis
• ...

Map Plotting Demonstration

A demonstration of map plotting is shown in the examples. In particular, we demonstrate extending external Google Maps into the notebook and the utility of the Geojson package to produce geographic plots, with markers and annotations that are directly linked into the plot.

- Example of extending Google Maps plots
- Example of Geojson plots that are linked to the Earth, e.g. regions on the kites used in the solar array selection example.
Offloading a processing pipeline to Amazon Cloud

Offload this pipeline to Amazon

Results

amazon = True
Cloud Enables Big Data Discovery on Phones

Plate Boundary Observatory Data

USGS Groundwater Data

NASA Kepler Data
Thanks!

Questions?

@vpankratius
pankrat@mit.edu