

Complex Data Explorer (CODEX) – A multi-use Machine Learning Powered Tool for Rapid Data Exploration

Jack Lightholder, Lukas Mandrake, Josh Rodriguez, Rob Tapella, Patrick Kage
jack.a.lightholder@jpl.nasa.gov

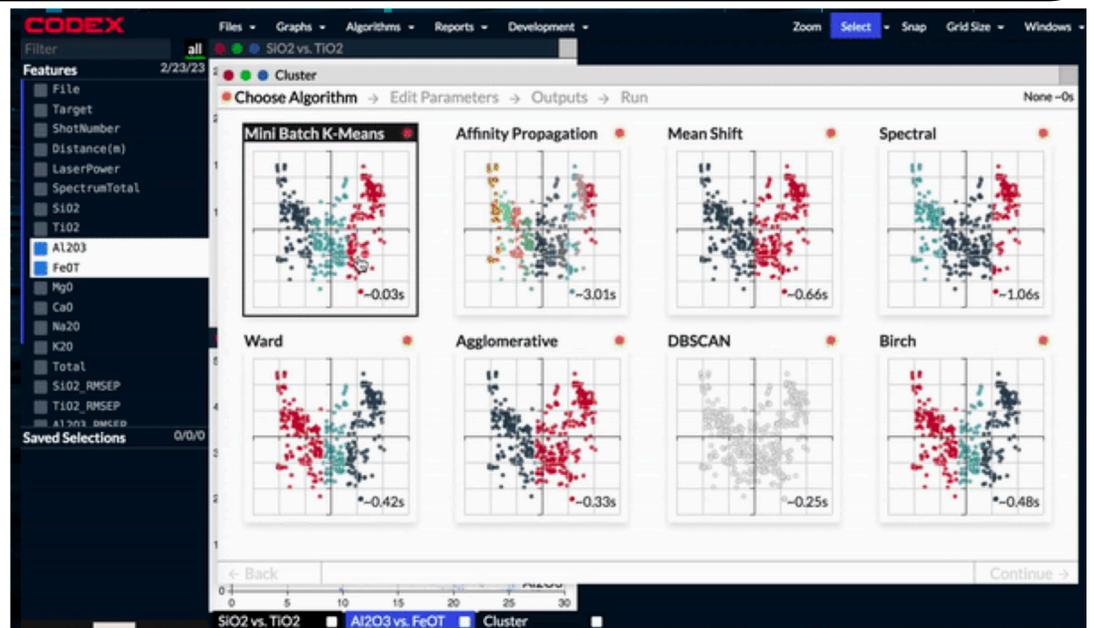
Abstract

Modern science datasets from missions like OCO-2 and telemetry records in Ops may have 500+ simultaneous measurements at each of millions of time samples. Scientists would often like to look through the record and discover not only expected trends but ones they did not initially guess, while Ops personnel perform the same task under serious time pressure should an anomaly occur. In both cases, the optimal environment for this rapid exploration large data would be one where visualizations were clear, interactive, and responsive, permitting the investigator to “play” with the data and gain rapid insight, falsify hypothesis, and make discoveries. Machine Learning (ML) has proven invaluable in providing some of these key data insights, but to do so in a statistically robust and reliable manner requires a data science professional and a lot of custom Python code, losing any sense of interaction and play. CODEX will address these concerns by providing a desktop-like environment with standard scientific graph types that are robust to rapid, powerful exploration.

A Common Need

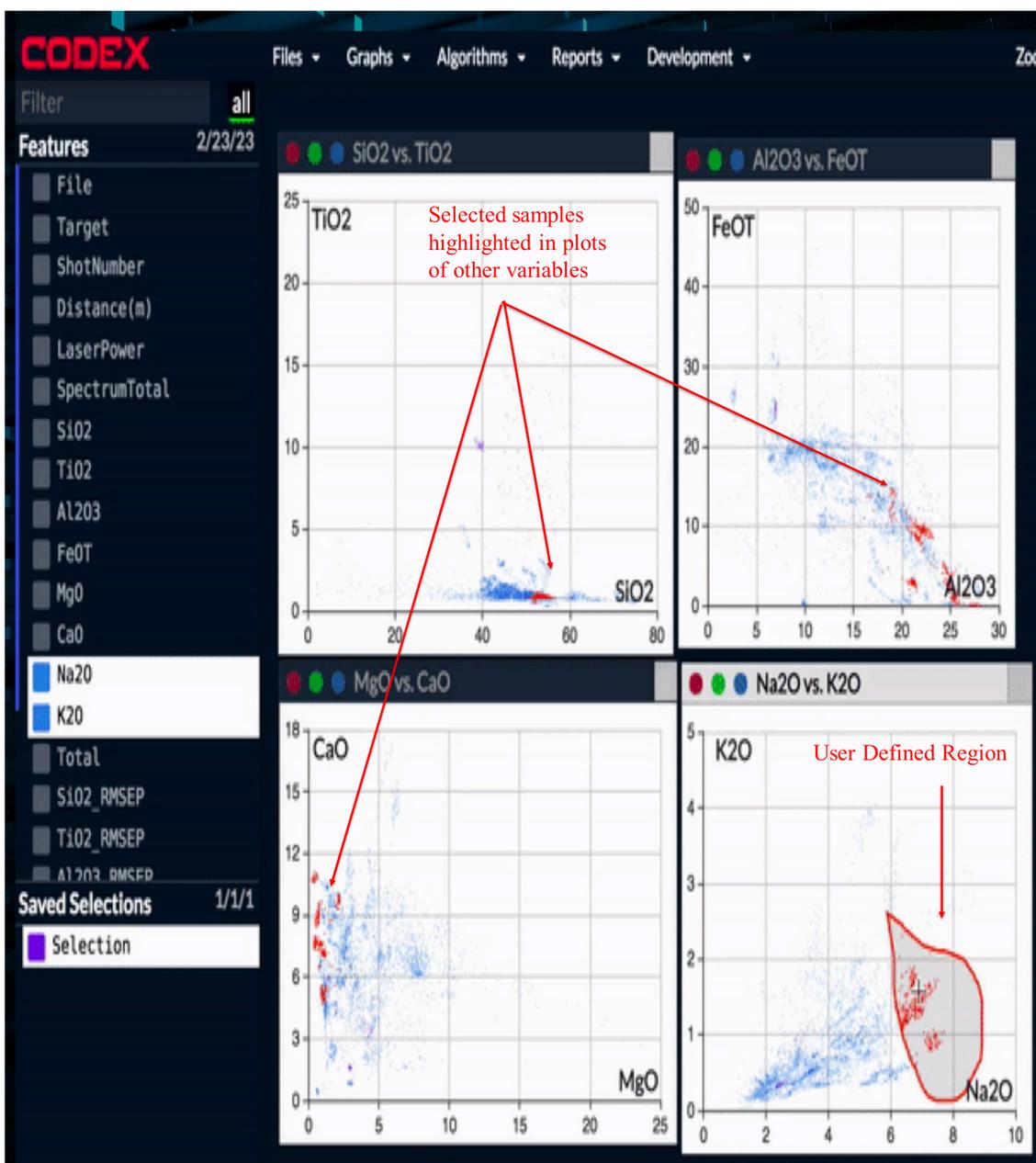
- High Dimensional time-series data
- Identify strange, outlier, or invalid values
- Interactively explore data
- Build/falsify hypotheses
- Interrogate relationships between cols
- Find more events like this
- Provide simple recipe to recognize events
- Create predictive, explanatory models
- How many families of data are present?

Machine Learning was made for this!



Intuitive exploration of the impact of algorithms, and their parameters settings, on your data. Rapidly explore algorithm categories to see which meet the needs of your application, without writing custom code every time!

Fast Hypothesis Testing



Intuitive exploration of multiple data components. Draw plot regions to see where those data points are in other aspects of your data for rapid intuition building.

Guiding Principles

- **Easy, interactive graphing**
 - scatter, heatmap, histogram, line, bar
 - linked: data here can be found everywhere else
- **Fast interactivity**
 - Humans learn best by manipulating playing
 - Slow batch analyses lose context & attention...
- **Continuous, visual guidance**
 - No question without rich support to guide answer
 - Permit visual selection & previews
- **Never stop working**
 - Long analyses run in background
 - Always forewarn of time & memory

Capabilities

- Clustering
- Regression
- Classification
- Dimensionality Reduction
- Feature Selection
 - Endmember Analysis
- Anomaly Finding
- Interactive / Linked Graphing

Conclusions

- Fast discovery of data issues & problems
- Fast intuition building
- Powerful ML techniques made visual
- Guidance for every step of exploration
- Doesn't replace Python or MATLAB
- Does start you off ready to do great work

Future Infusion & Applications

- Mission operations data – build quick intuition about vehicle health and conduct on the fly trending.
- Science data – explore data sets to locate areas of interest for deeper study.
- Complex data interactions – leverage machine learning to gain insights to data characteristics which separate populations of data.