GSAW 2020

Jack Lightholder
Lukas Mandrake
Joshua Rodriguez
Rob Tapella

CODEX

Complex Data EXplorer

A Common Need

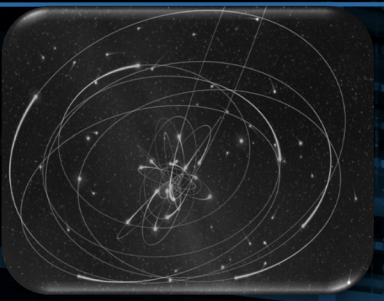
Dr. Lukas Mandrake, Jack Lightholder

A Common Need

Model generates 1M potential mission trajectories w/ dozens of fitness metrics

- Do the trajectories fall into families with similar behaviors?
- How many such families are there?
- What are the uniquely identifying characteristics of each family?

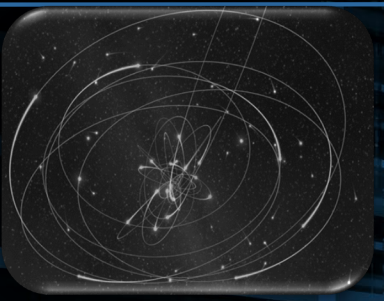
A Common Need



New OCO-2 data release. 10M soundings. Validation team needs to sign off.

- Does this version of the retrieval code match past behaviors?
- If not, where & when do they differ?
- What key atmospheric conditions correlate with the mismatch?

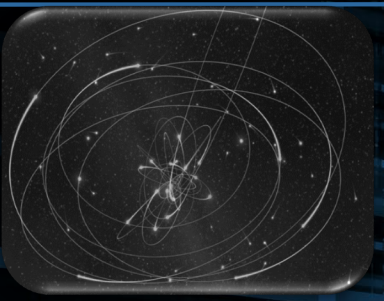
A Common Need



Hardware fault just occurred onboard! Need to find root cause ASAP.

- Does the pre-fault telemetry show any differences vs. the past?
- If so, what are the key telemetry channels or science products to examine?
- What times in the past looked similar to the pre-fault state?
- Rapid hypothesis forming, testing, falsification

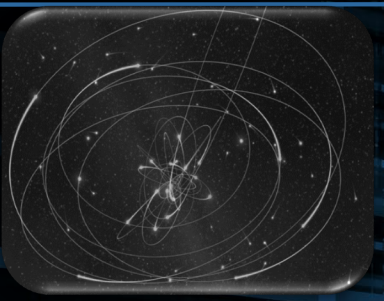
A Common Need

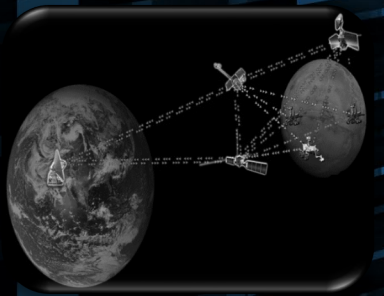


Signal lock to an active mission was just lost unexpectedly… again

- How unusual is this event given the past mission experience?
- What were the most unusual readings in signal metrics before loss?
- Are there trends in losing lock? What do they depend on?

A Common Need

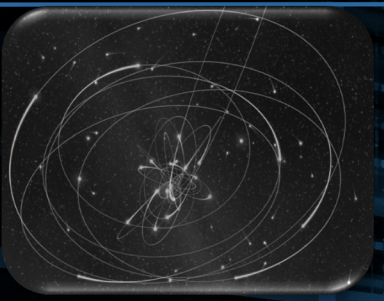


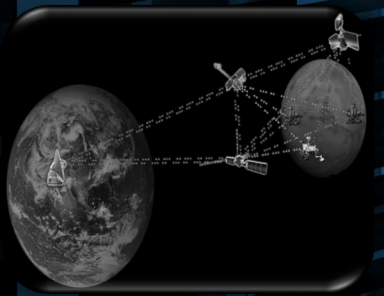


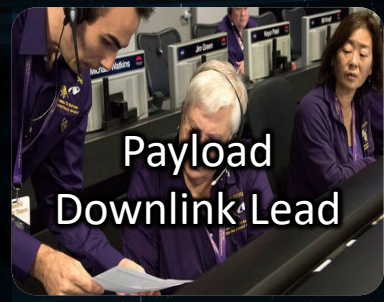
New science data just arrived… is it high priority?

- Compare to past findings… how unusual is it?
- Compare to targets of interest… is it similar?
- Do the expected correlations and relationships hold?

A Common Need



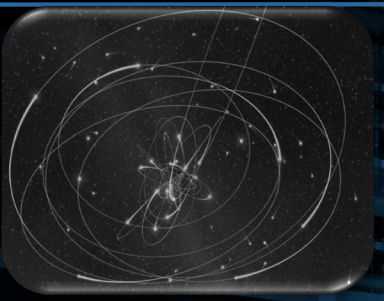


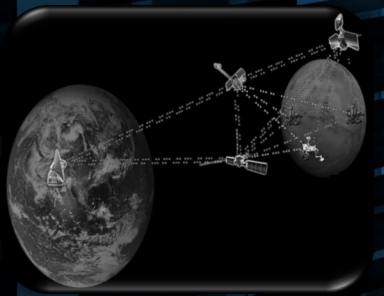


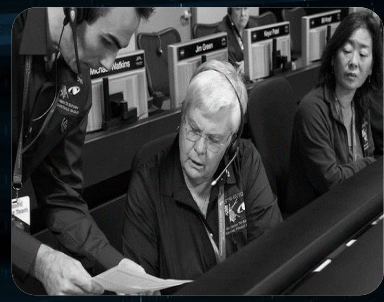
Regular, prolific telemetry arrives daily

- Any likely problematic values?
- Are co-varying trends as expected?
- If oddities appear, which channels and times?

A Common Need



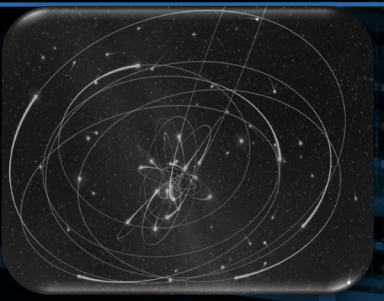


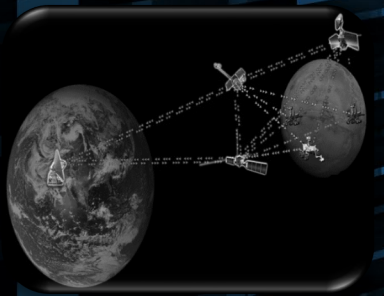


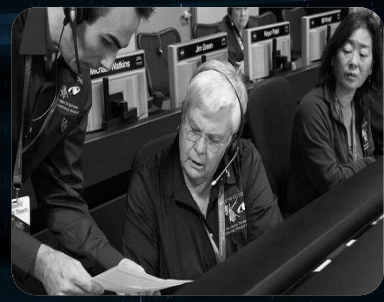
Interesting phenomenon found! But data is too big…

- Find more events like this… where, when?
- What's simplest "recipe" to find these events?
- Select them out of the data for later analysis

A Common Need



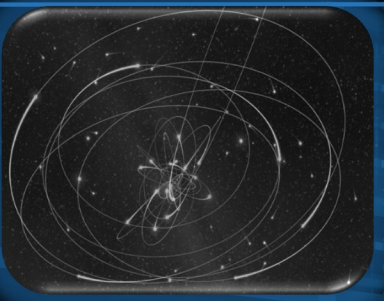


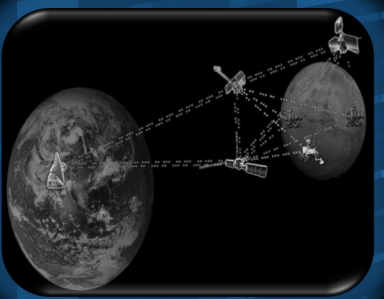


"Take a look at my data and see what you think."

- Any troublesome issues? High correlation? Sparsity?
- Perform quick modeling and make trial products
- Try out many technologies to find good matches
- Sanity check expected behavior & provide feedback

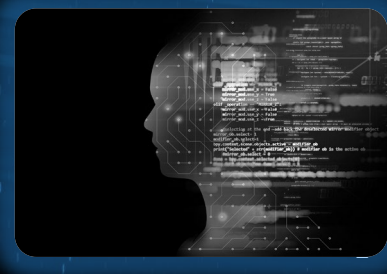
A Common Need





- 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?

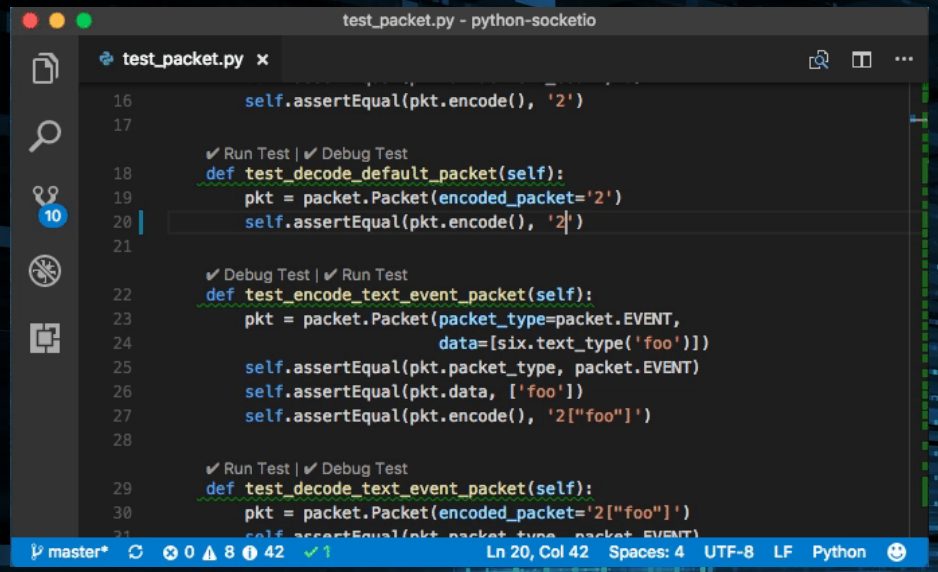
Machine Learning was made for this!

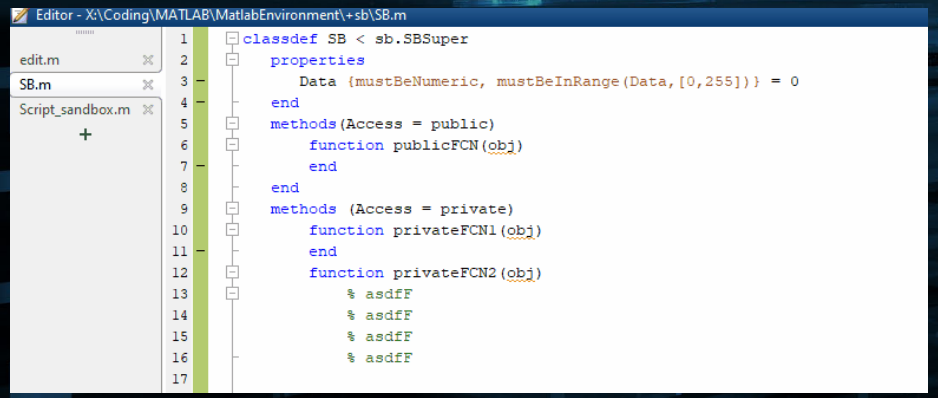


Current State of Art

Dr. Lukas Mandrake, Jack Lightholder

Current State of the Art

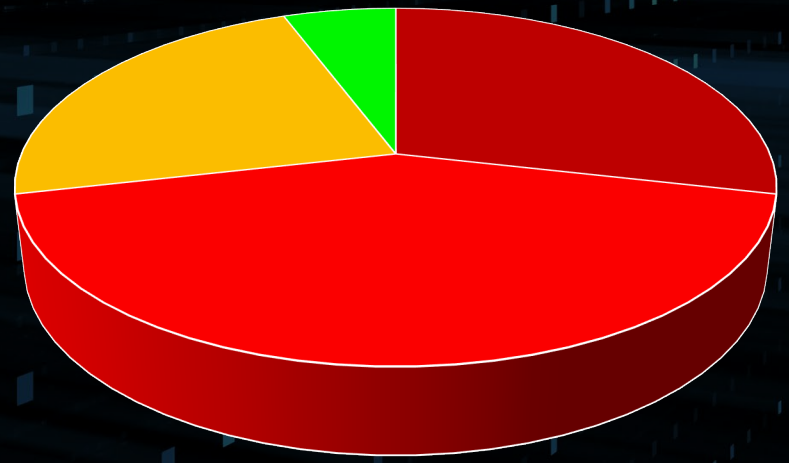




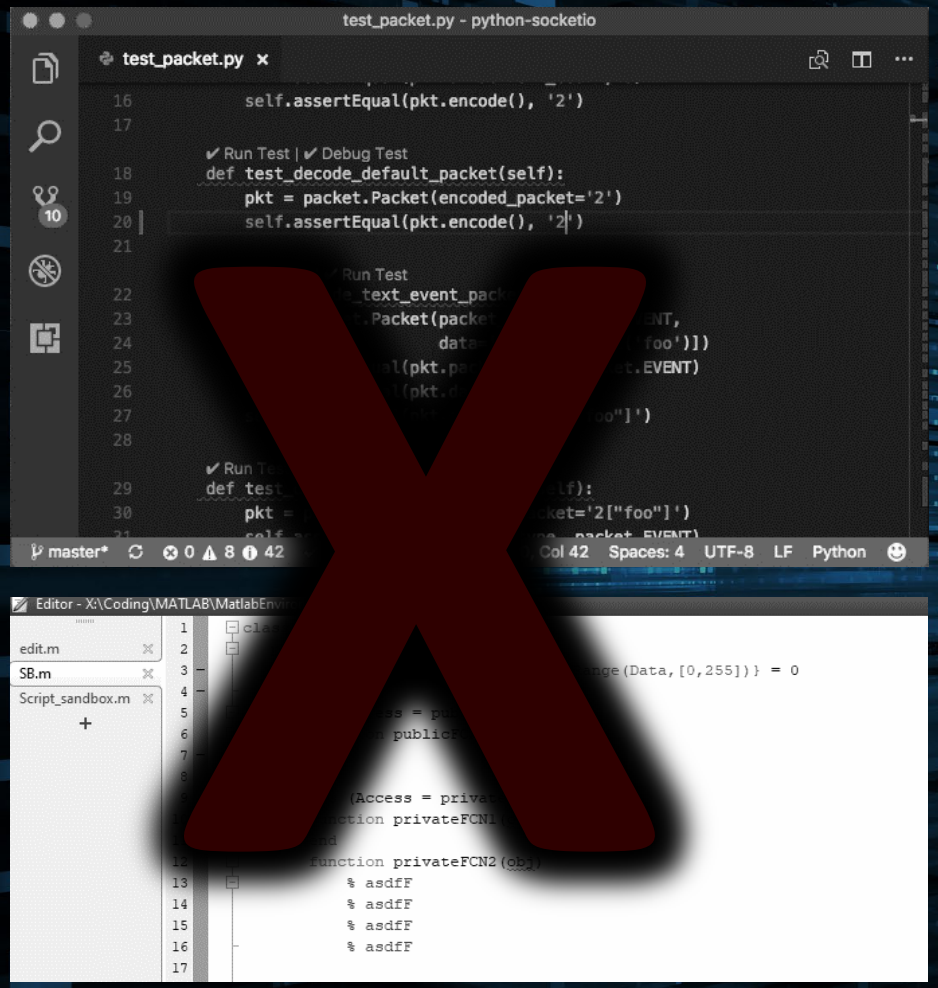
- Python, Matlab, R scripts… lots of them
- Great for routine processing / checking known issues
- Very inefficient for exploration
  - Did you search for all potential problems?
  - Did you verify all assumptions?
  - Did you thoroughly visualize your entire data?

Of Course not!

Custom Code for Data Exploration

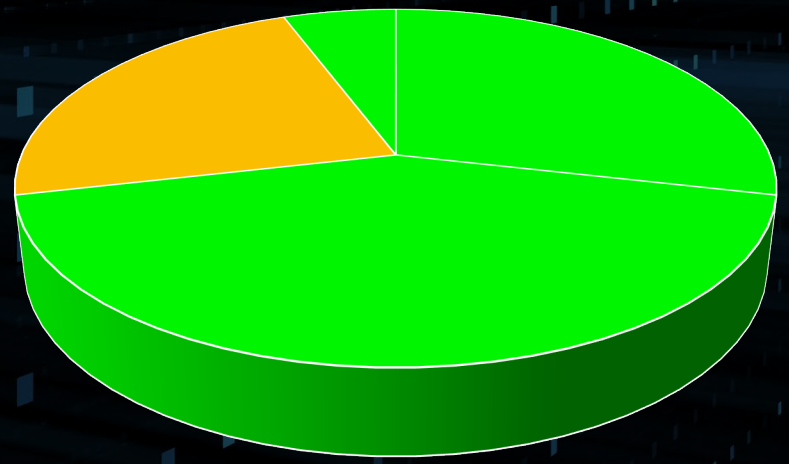


New Concept



- Don’t fuss over buggy code
- Interactively graph and explore
- Beautiful interface to many ML approaches
- Tools to address the Common Need
- Writes the code for you afterwards!

Custom Code for Data Exploration

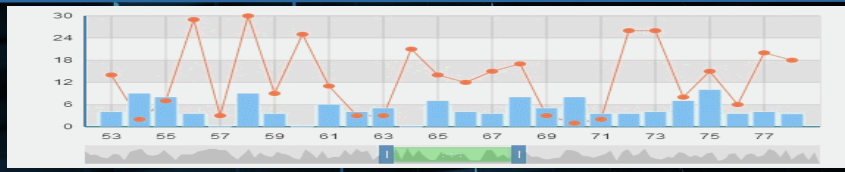


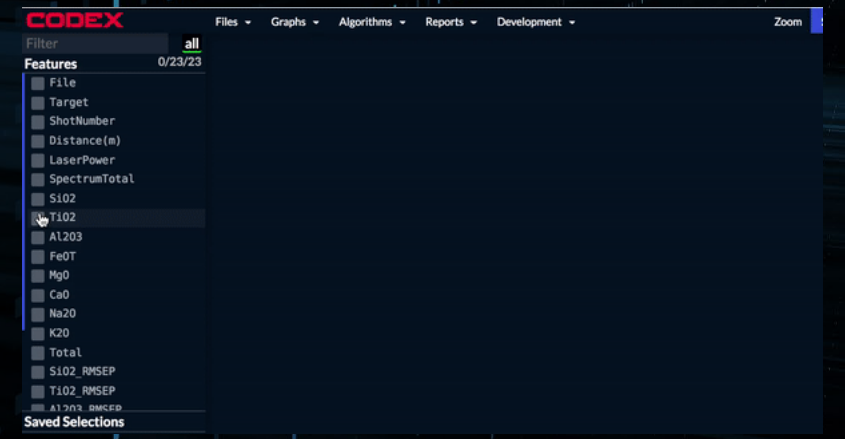
CODEX

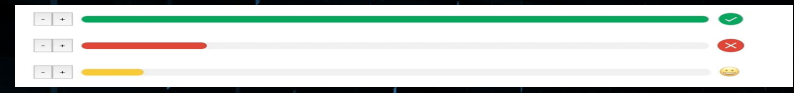
Dr. Lukas Mandrake, Jack Lightholder

Guiding Principles

- Easy, interactive graphing
  - scatter, heatmap, histogram, line, bar
  - linked: data here can be found everywhere else
- Continuous, visual guidance
  - No question without rich support to guide answer
  - No obscure numbers; permit visual selection & previews
- Fast interactivity
  - Humans learn best by manipulating and studying: playing
  - Slow batch analyses lose context & attention…
- Never stop working
  - Long analyses run in background
  - Always foreworn of time & memory needs for all choices







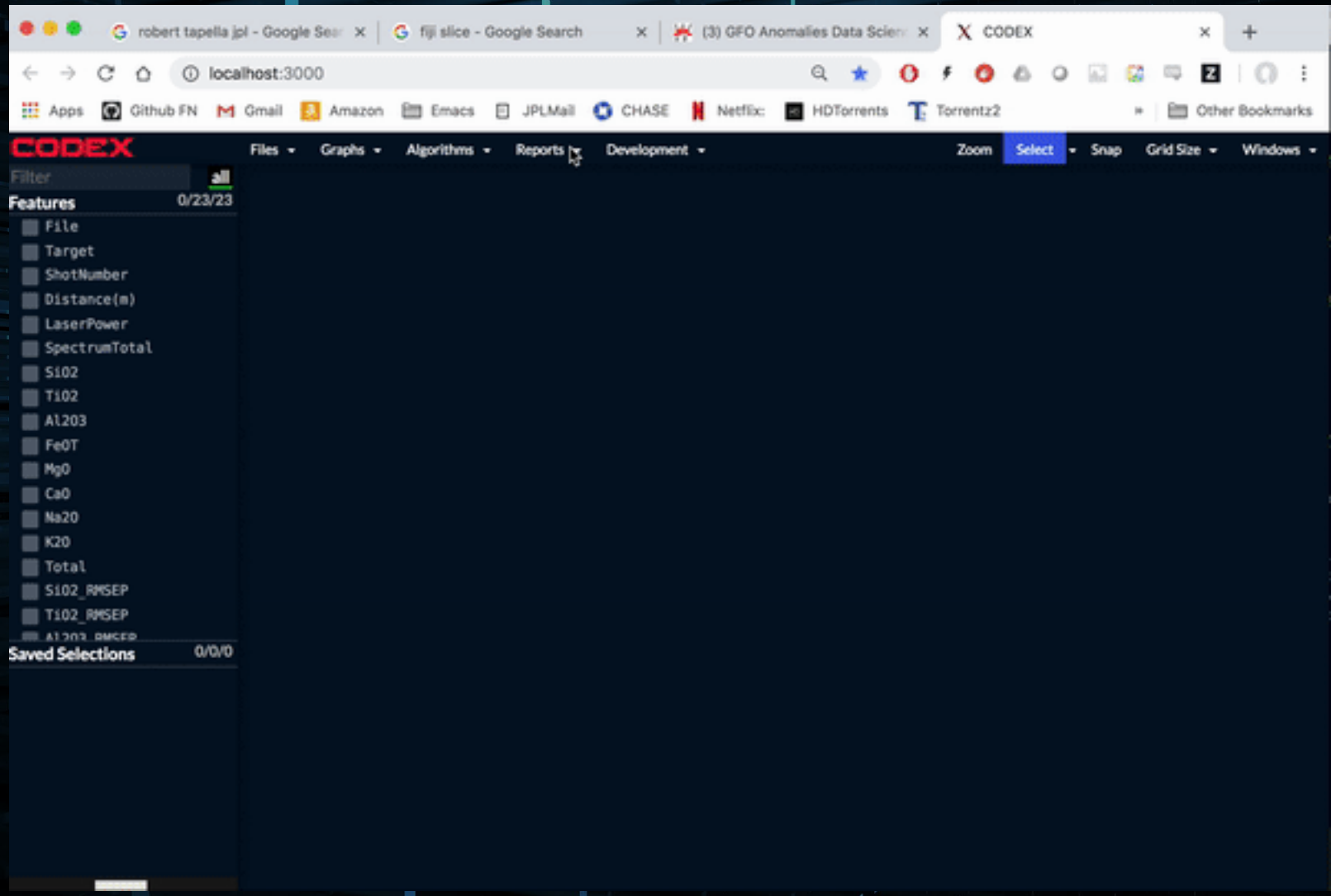
Initial Data Scan

- Easy: NaN's and Inf's
- Moderate: Sigma-outliers
- Subtle: Repeated values

- Which rows/columns affected?

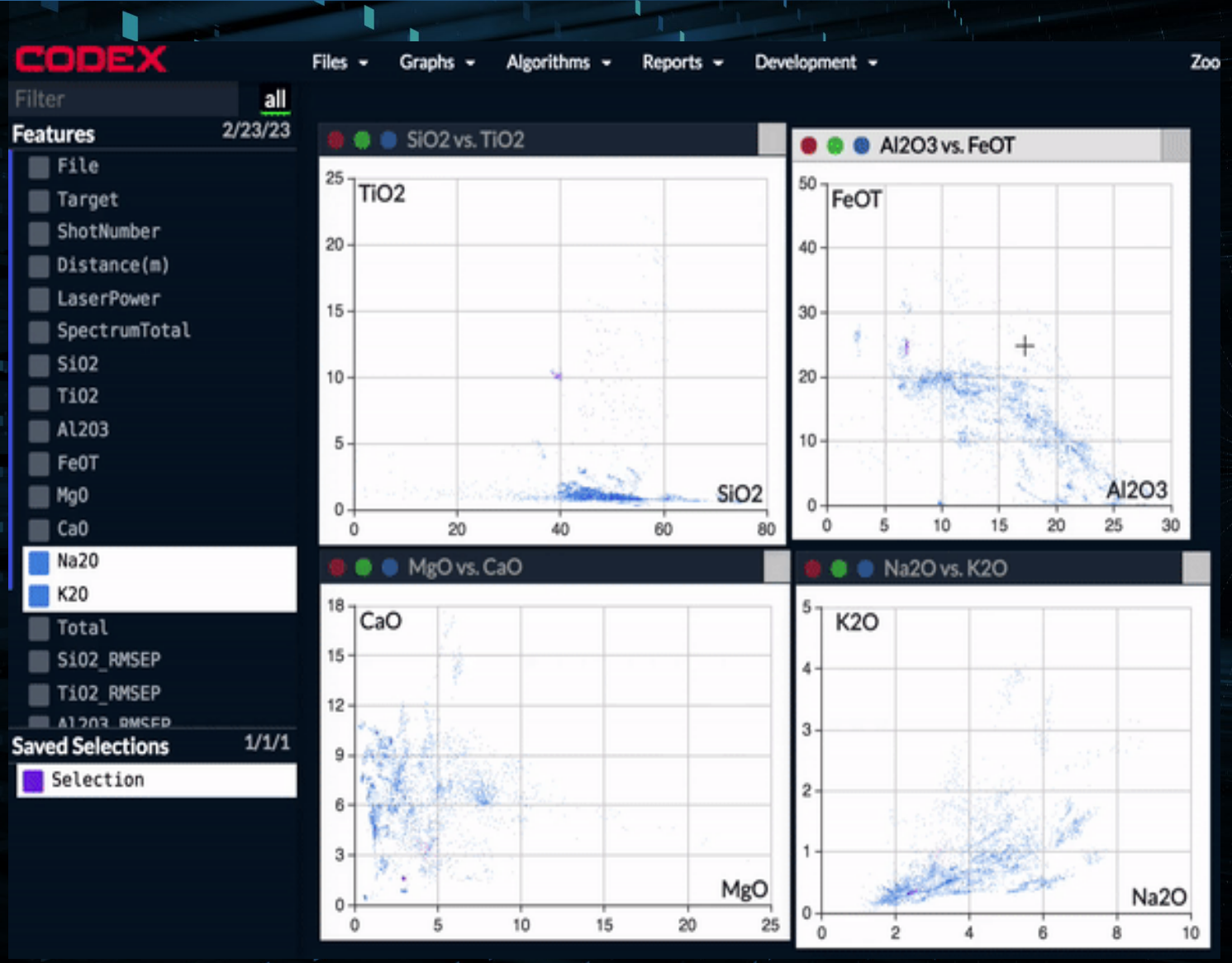
- Just some bad rows to filter? Bad columns?

- "Sweep" thresholds to observe effects



Interactive, Linked Graphs

- Make discoveries just by selecting data in a graph
- “I wonder what this group of data means?”
- Rapidly prove / falsify hypotheses
- Simple concept / “Wow!” factor to researchers
- Perfect example of discovery through play
- Will be extended to all graph types



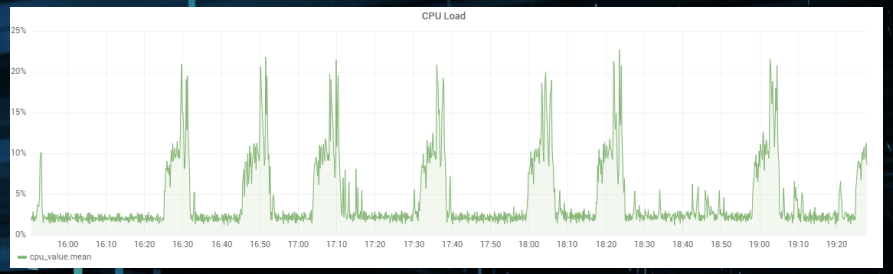
How Many Kinds? (Clustering)

- Can’t plot everything vs. everything else…
- Automatic search for interesting groups
- Perfect example of visual guidance
- Focus-of-attention tool

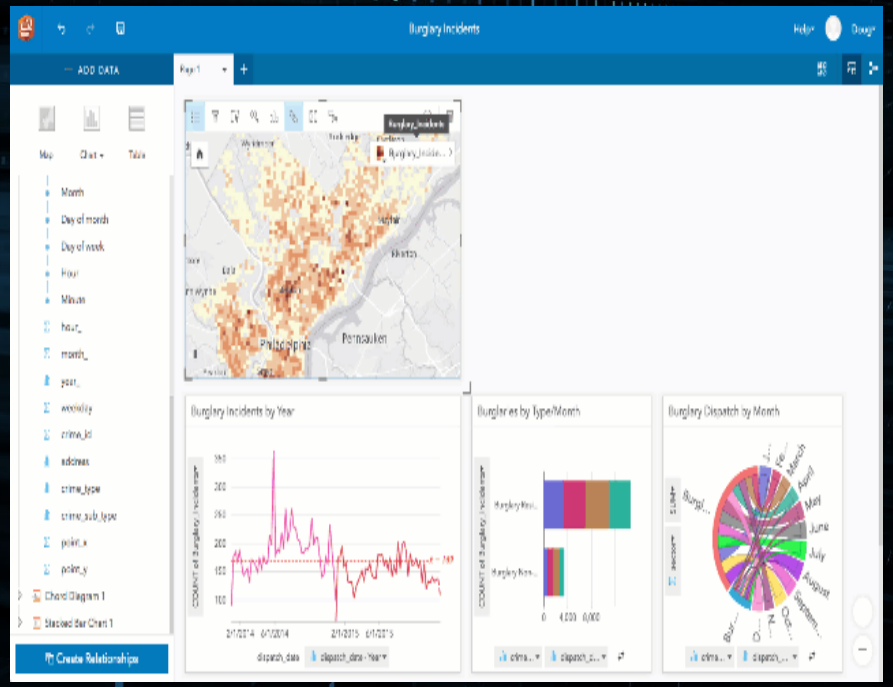
Find More Like This (Interactive Classification)

- Starts with examples of desired events
- Brush them in time: positive label
- Brush undesired regions: negative label
- Classify remainder of time as (un)desired
- Interactively modify “mistakes” to refine
- Finish with potential events & trained model

(notional, not yet implemented)

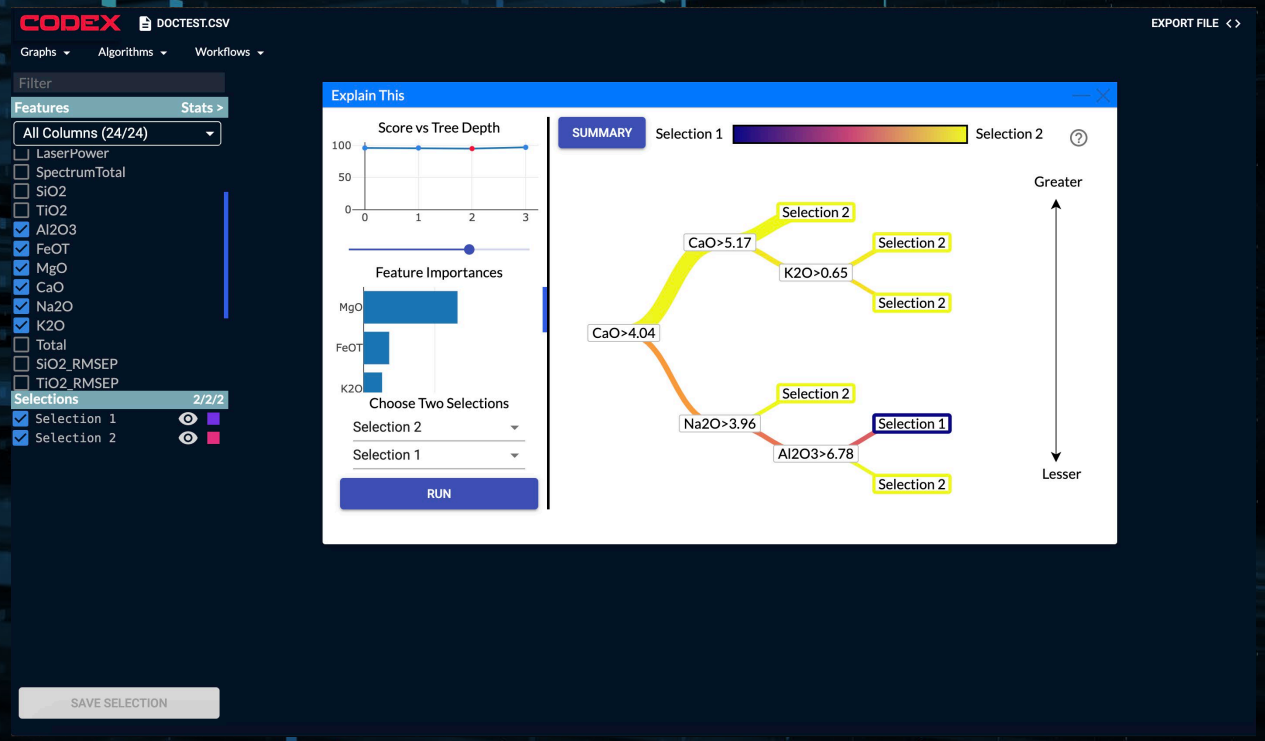


(notional, not yet implemented)



Explain This (Feature Selection / Endmembers)

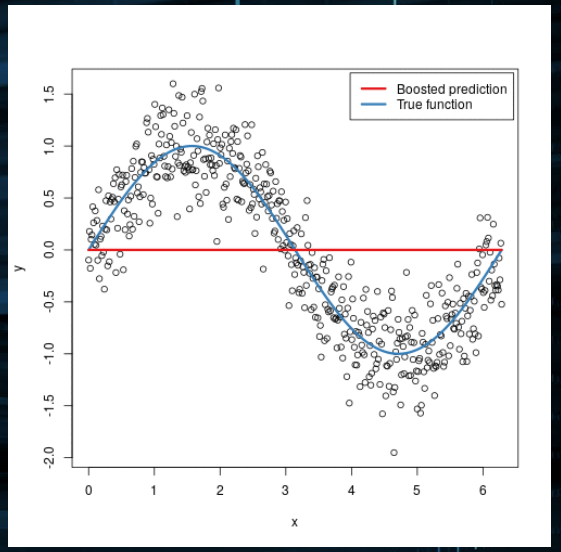
- Starts with subsets of interest
  - clustering or user-selected
- What columns best explain what's special about one group vs another?
- If all data in this subset were made up of a linear mixture of N samples, which N best explain the data?
- For each sample, what mixture of these endmembers is required?



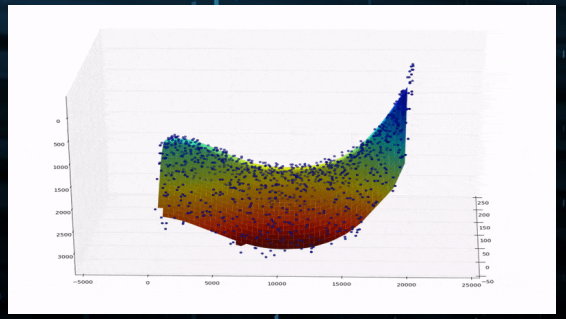
Predict This (Regression)

- Predict one column using others
- Simple: multivariate curve fitting
- Complex: Model-free ML regression
- Which columns proved most useful?
- What % variance explained for each?

(notional, not yet implemented)

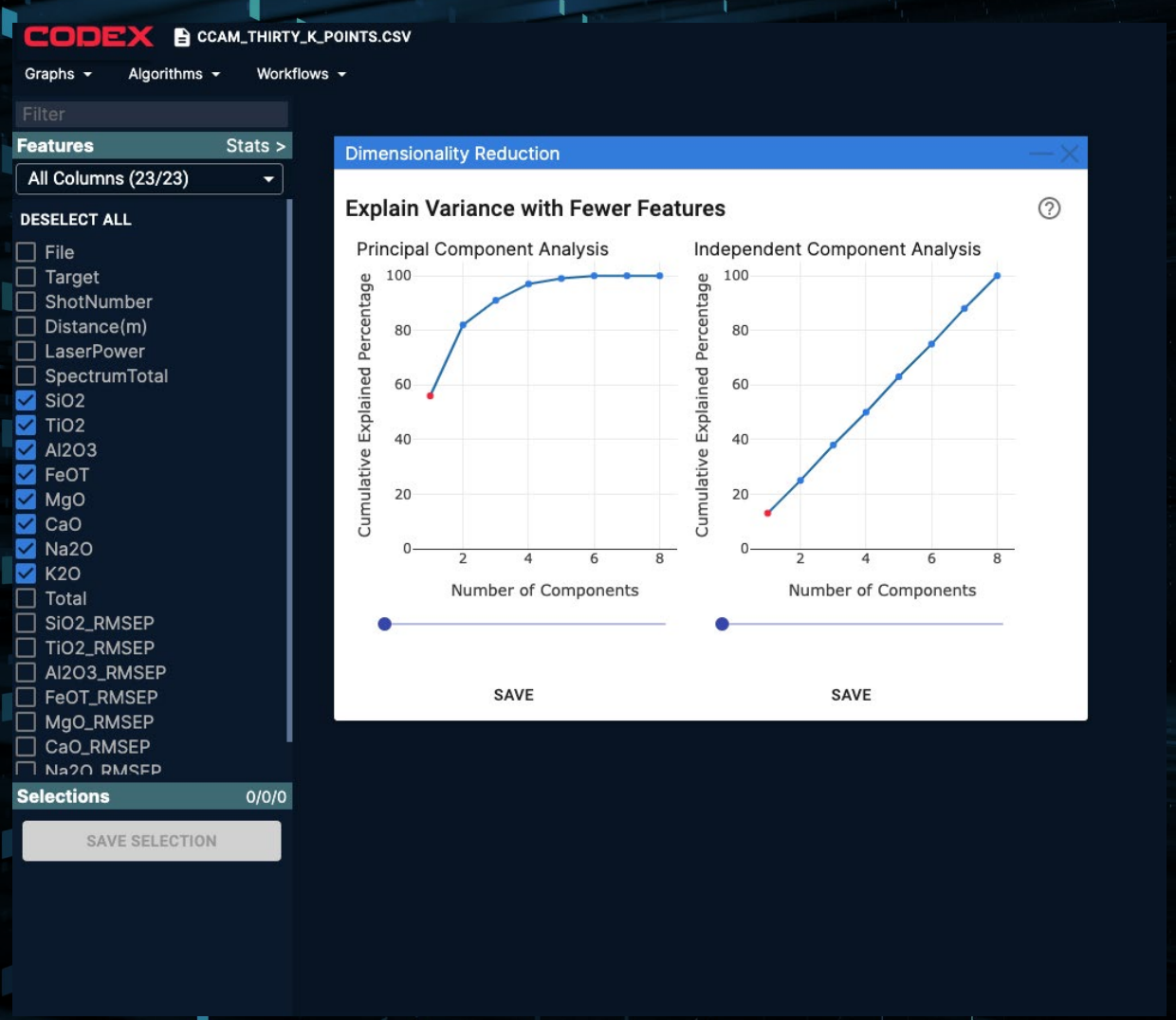


(notional, not yet implemented)



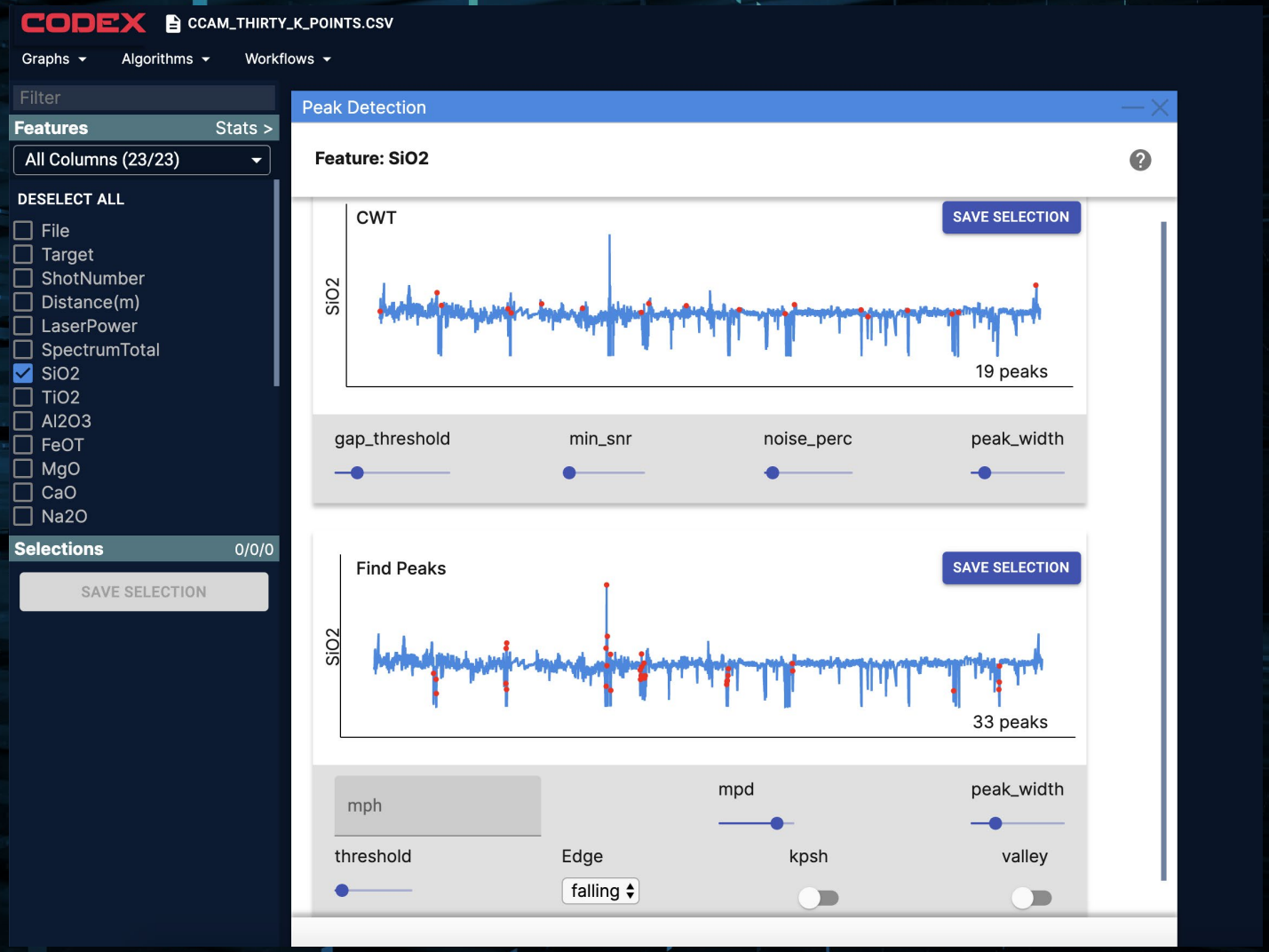
Remove Correlations (Dim Reduction)

- Never need all the columns
- Because Physics exists, there are relationships between
- Find smaller set of new columns
- Can reveal essential relationships between original columns
- Graphs even hard-to-visualize datasets



Find Strange Things (Anomaly Finding)

- Starts with “normal” times
- Detect likelihood of normalcy elsewhere
- Least normal regions are anomalies
- Focus of attention for further investigation

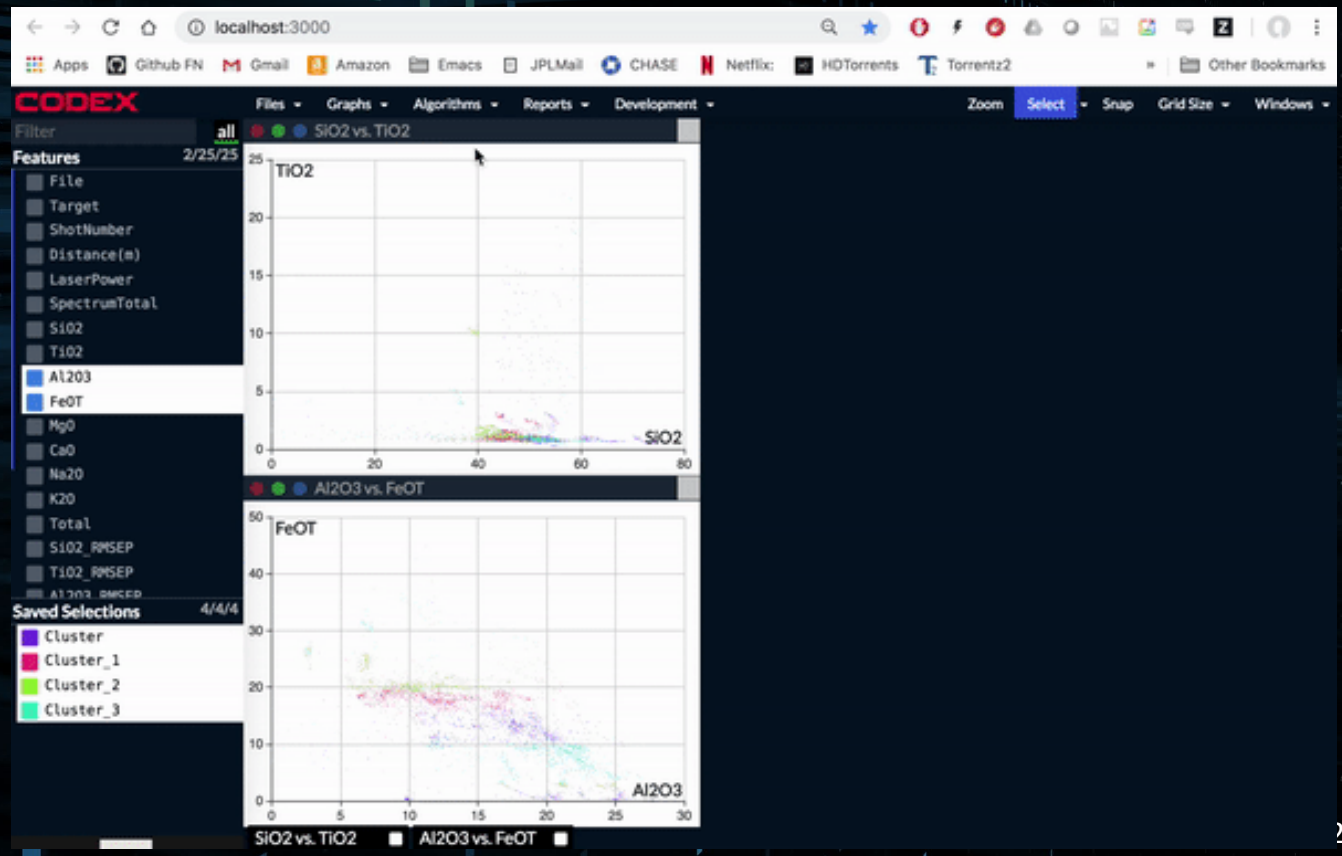


Don’t just explore: Get a Coding Head-Start!

- Everything user does is generating Python code!
- Can just “export” code to a file at the end
- User just picks up in Python and continues deeper analysis
- Unprecedented support & time savings

```
def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodename()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print ' %s [label="%s' % (nodename, label),
    if isinstance(ast[1], str):
        if ast[1].strip():
            print '= %s"];' % ast[1]
        else:
            print '"]'
    else:
        print '"];'
        children = []
        for n, child in enumerate(ast[1:]):
            children.append(dotwrite(child))
        print ' %s -> (' % nodename,
        for name in children:
            print '%s' % name,
```



CODEX: Know Thy Data

- 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

Follow our progress on GitHub:
https://github.com/NASA-AMMOS/CODEX