I'm not your boyfriend!
You totally are.
I'm casually dating a number of people.
But you spend twice as much time with me as with anyone else. I'm a clear outlier.
Your math is irrefutable.
Face it—I'm your statistically significant other.

... okay, but because you said that, we're breaking up.

https://xkcd.com/539
### LEAST FREQUENCY OF OCCURRENCE (LFO)

<table>
<thead>
<tr>
<th>filename</th>
<th>cnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>R:61d8vk4iYPSYJFSB.txt</td>
<td>1</td>
</tr>
<tr>
<td>R:AGj7CSEPG.pdf</td>
<td>1</td>
</tr>
<tr>
<td>R:ALJEI1110:1U8GTHS65.pdf</td>
<td>1</td>
</tr>
<tr>
<td>R:BXD6959:W1:3K6157.zpl</td>
<td>2</td>
</tr>
<tr>
<td>R:GN3GP41:MW66W6QY.jpg</td>
<td>2</td>
</tr>
<tr>
<td>R:ENH0037:T7Z9JE1R.doc</td>
<td>2</td>
</tr>
<tr>
<td>R:ACKB1441:0UYHJKUNP.doc</td>
<td>2</td>
</tr>
<tr>
<td>R:YBG0384:PF4304LX.doc</td>
<td>2</td>
</tr>
</tbody>
</table>

**Least-Commonly Accessed Files**

LEAST FREQUENCY OF OCCURRENCE SAMPLE CODE

SQL
SELECT cs_user_agent, count(*) as cnt FROM proxy WHERE [...] GROUP BY cs_user_agent ORDER BY cnt ASC LIMIT 100

Splunk
index=proxy method=POST status=200 | stats count by cs_user_agent | sort +count | head 100

Python
# Using the ‘pandas’ module and DataFrames
print proxy_df.value_counts(by="cs_user_agent", sort=True, ascending=True).head(100)
“I’ve never seen this service on any of my 100,000+ systems before. That’s pretty suspicious.”
WHEN IN DOUBT, VISUALIZE!

http://skepdic.com/graphics/elvistoast.jpg
SCATTER PLOTS

PCR Shift (Data Exfiltration)

Our exfil host

Dirty PCR vs Baseline PCR chart.
**SIMPLIFIED SCATTER PLOT CODE**

```python
trace = go.Scattergl(
    x=points_x,
    y=points_y,
    mode='markers'
)

data = [trace]

pyo.iplot(dict(data=data))
```

Basic scatter plot: <= 6 LOC
"Most users don’t change upload habits often, so benign activity should fall close to the ‘no change’ trend line."

SCATTER PLOT USE CASES

**Strong**
- Comparing numerical data on two axes
- Malicious data violates a clear correlation
- You need to establish whether a correlation even exists

**Weak**
- Working with non-numeric data
- The variables on the two axes have weak/no correlation
- Need to compare on more than two axes
Command Line Lengths for Shell Processes
SIMPLIFIED BOX PLOT CODE

data = [  
go.Box(  
y=shells['command_line_length'],  
    boxpoints='all'  
  )  
]  

layout = go.Layout(  
    title="Command Line Lengths for Shell Processes"  
)  

pyo.iplot(dict(data=data, layout=layout))
“I have no idea what’s normal for this column of numbers, let alone whether there are any outliers.”
You have to be careful doing this. Sometimes, when you push the whisker down, dynamite explodes.
A form of unsupervised machine learning.

Iteratively split dataset on random dimensions and their values until you can’t split anymore. This is a “tree”.

Do this several times to grow the tree into a “forest”.

Average depth across all trees for each point reflects “outlierness”.

---

(a) Isolating $x_i$

(b) Isolating $x_o$
ISOLATION FORESTS SAMPLE CODE

https://github.com/DavidJ Bianco/Clearcut
ISOLATION FOREST USE CASES

**Strong**

- Medium or high dimensional numerical data, or things that can be converted (e.g. log file entries)
- Can outperform other clustering mechanisms (faster, lower memory requirements)
- You get to claim you use “machine learning”

**Weak**

- Feature engineering for converting non-numeric data can be challenging
- Difficult to visualize at high dimensions

“Which of these HTTP log entries are most unlike the others?”
When it comes to outlier detection, you’ve got options. Don’t be afraid to play around a bit and see what works best for you!
I HAVE A QUESTION AND/OR WISH TO MAKE A SHORT SPEECH...

DAVID J. BIANCO
@DavidJBianco
ThreatHunting.Net
Detect-respond.blogspot.com