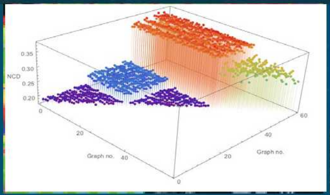
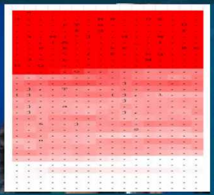
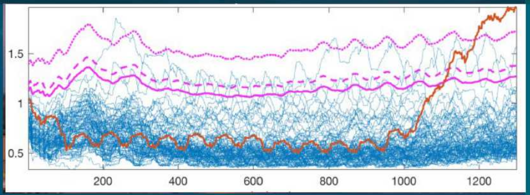
Sandia National Laboratories

SAND2019-2563PE

Compression Analytics for Data Science







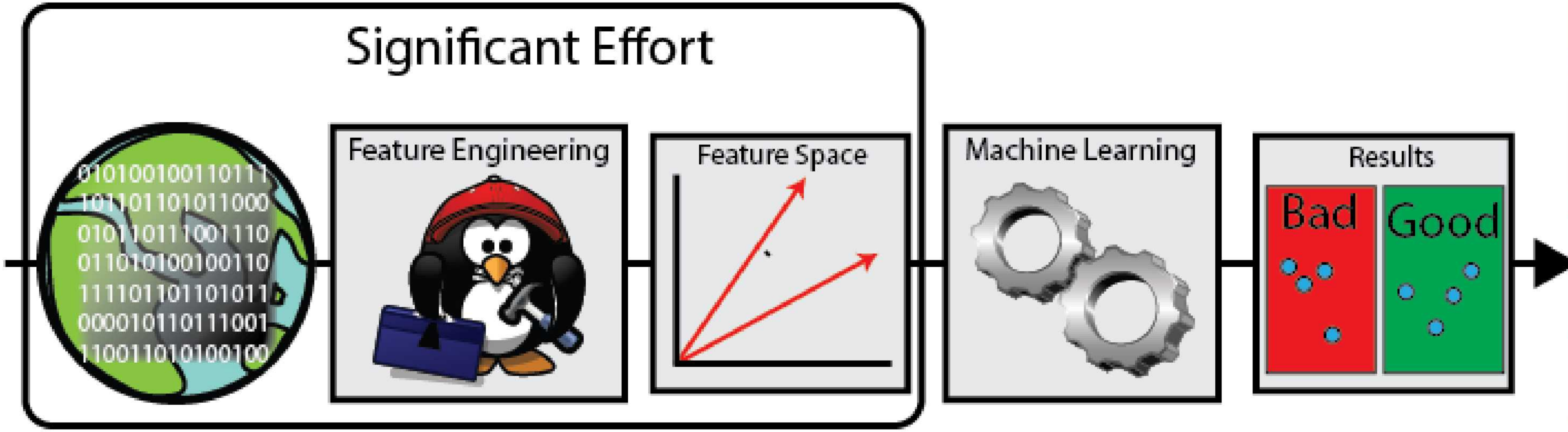
Travis Bauer, Christina Ting, Andrew Fisher,
Rich Field, Tu-Thach Quach, Tom Brounstein,
Alex Killian, Randy Wells

ENERGY NNSA

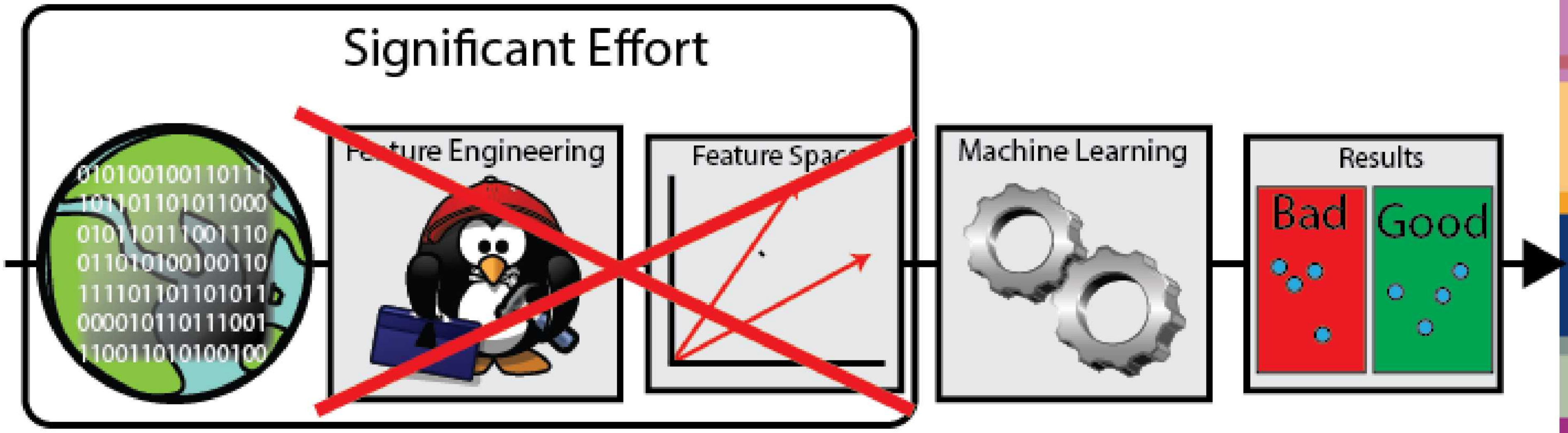
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Most machine learning techniques require specific types of data.

But the world doesn't come in the form of vectors and preprocessing data is a lot of work.

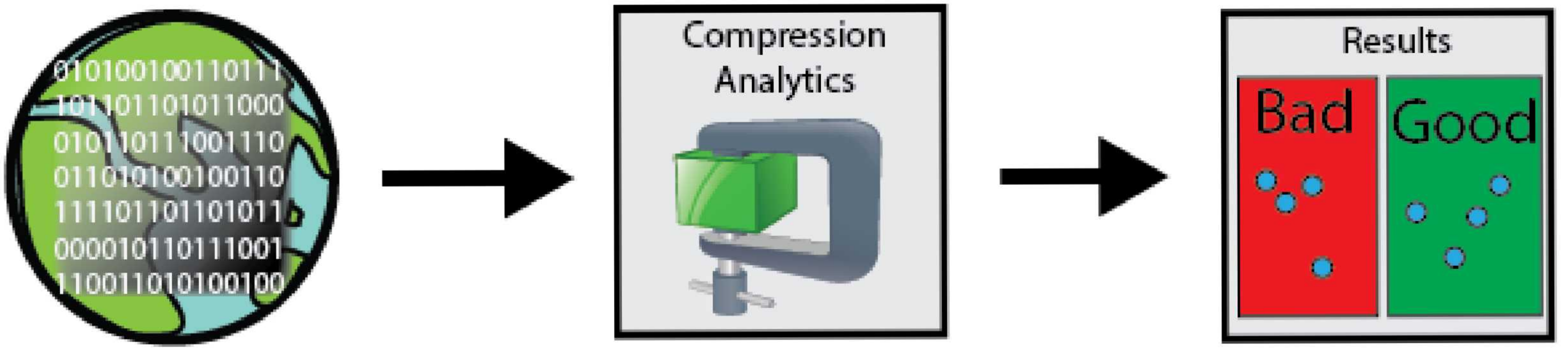


It would be nice if we could find a family of algorithms that let's us cut out the expensive feature engineering portion of our work.



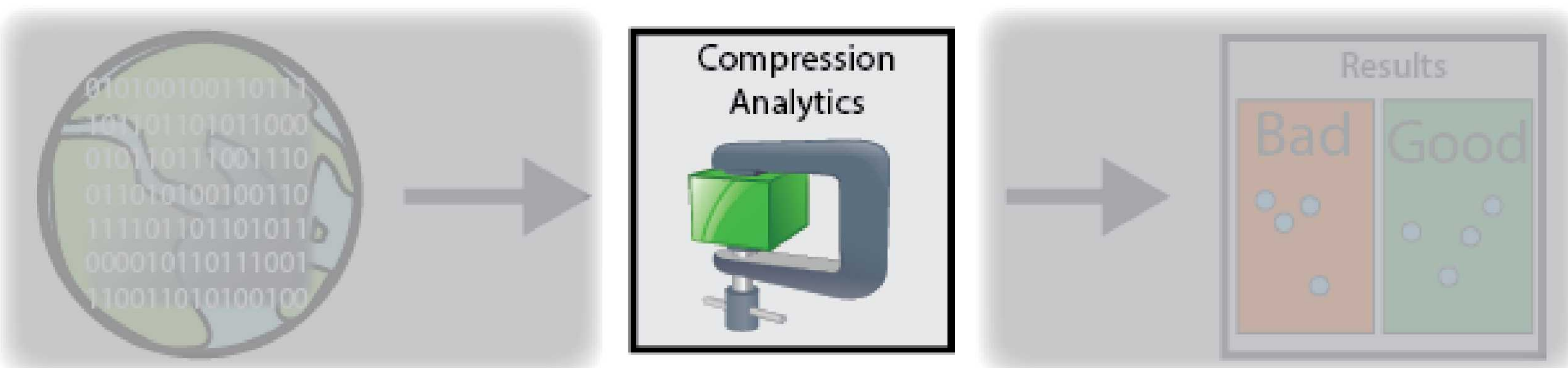
It would be nice if we could find a family of algorithms that let's us cut out the expensive feature engineering portion of our work.

Compression-Based Analytics provides capabilities that can be used this way.



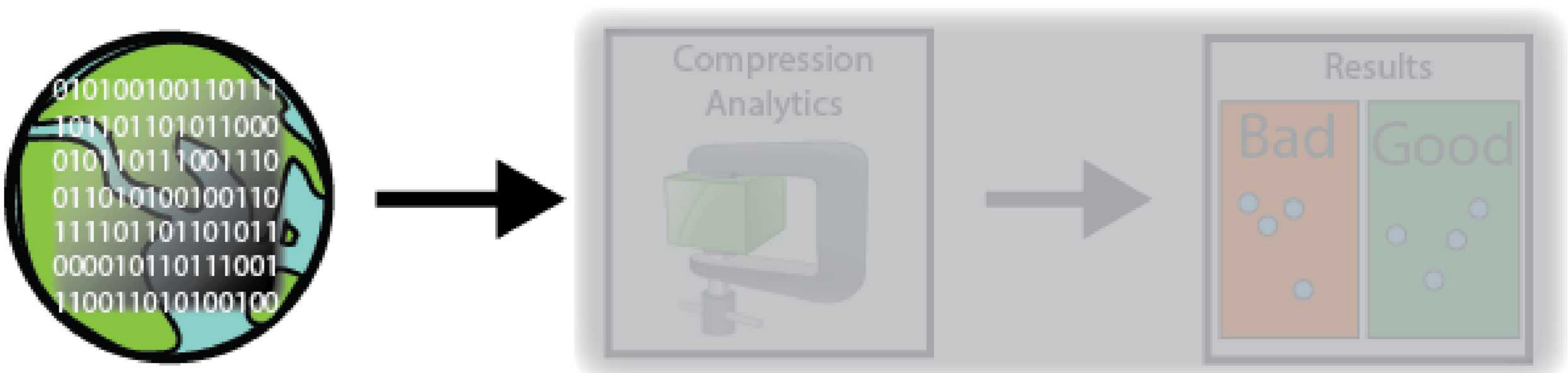
I’m going to show you how compression analytics works, where it works, and how we might make them work better

HOW compression analytics work

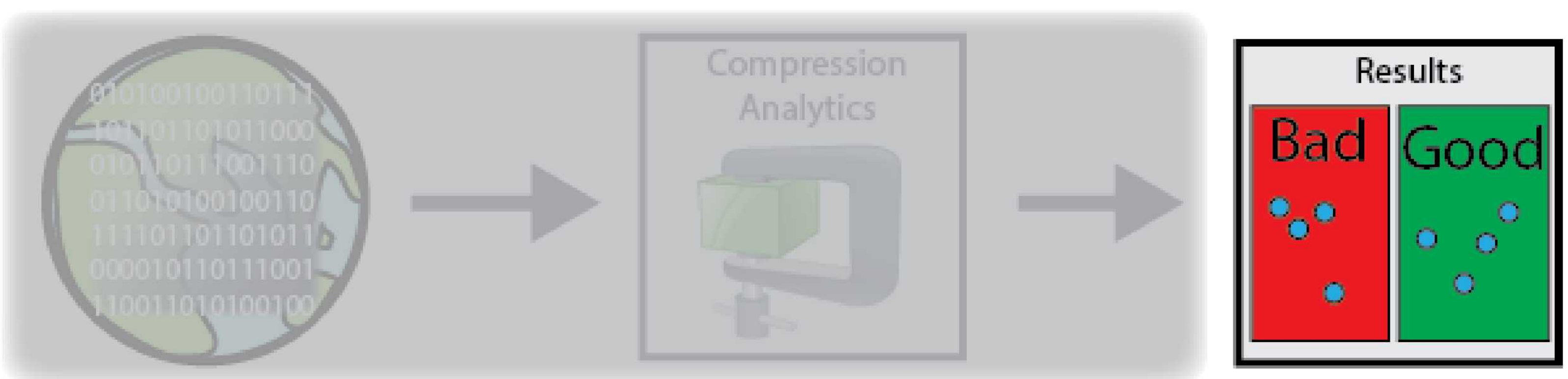


I'm going to show you how compression analytics works, where it works, and how we might make them work better

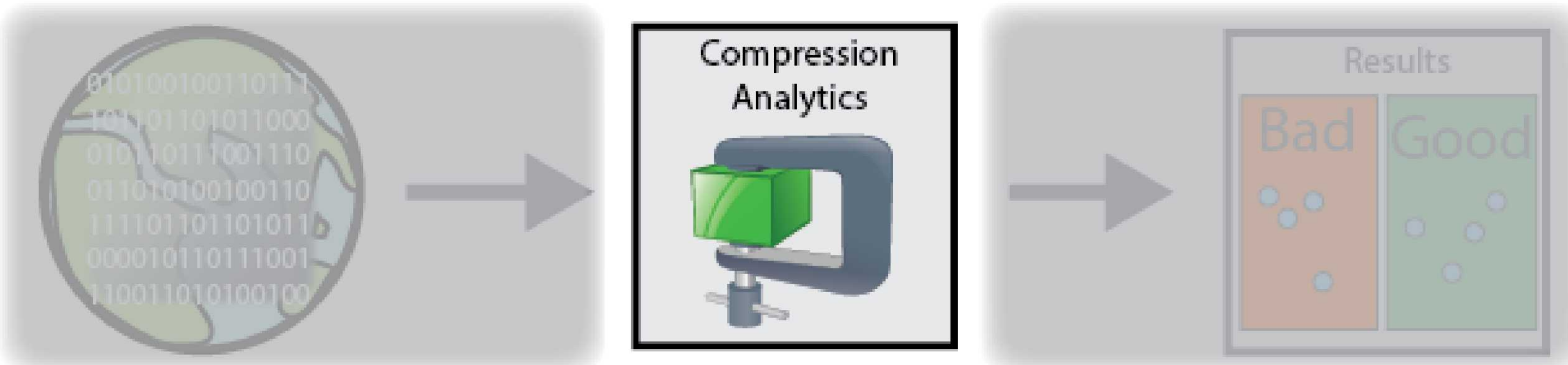
WHERE compression analytics work



How to make compression analytics work BETTER



HOW compression analytics work



Compression analytics work by approximating the amount of shared information in data.

Compression algorithms approximate the amount of information in data

A String of 10K ‘a’s

Much Data

Almost no Information

'a'*10000

10K Truly Random Characters

Much Data

Much Information

‘TNYSTCACWIIMLZX EWAGG…’

Compression is an upper bound on the amount of information.

A String of 10K 'a's
Much Data
Almost no Information

Length_compressed(
'a'*10000)

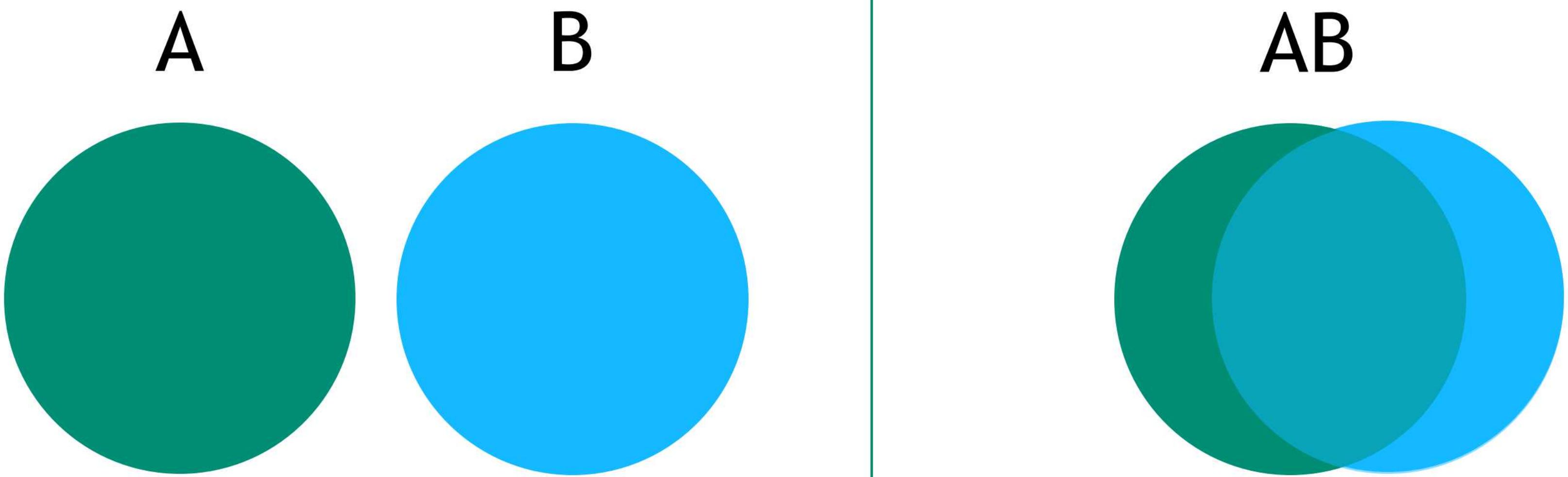
86 Bits

10K Truly Random Characters
Much Data
Much Information

Length_compressed(
'TNYSTCACWIIMLZXEWA
GG…')

68,418 Bits

We can compress two items together to approximate the amount of shared information.



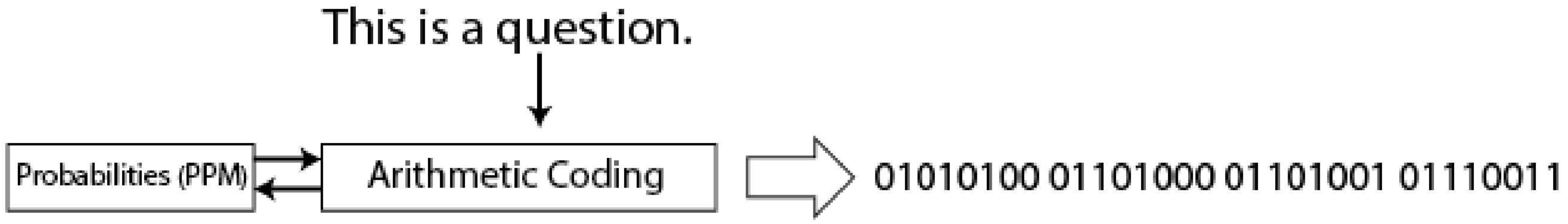
A and B are Completely Different

A and B Overlap

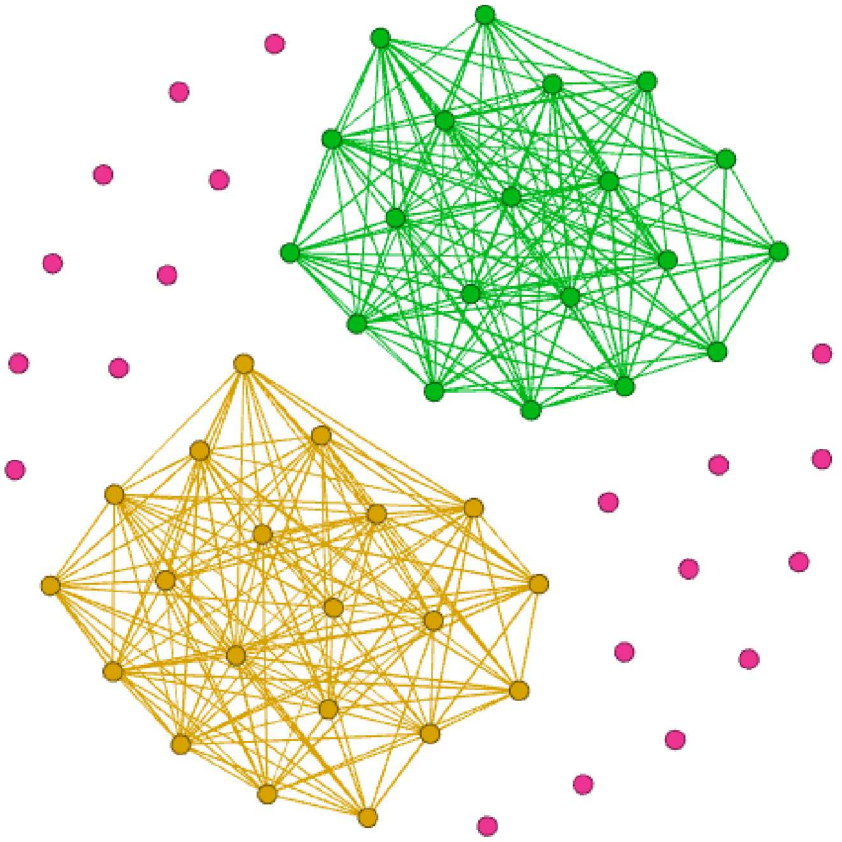
$$NCD_Z(A,B) = \frac{Z(AB) - \min\{Z(A), Z(B)\}}{\max\{Z(A), Z(B)\}}$$

Li, Ming, et al. "The similarity metric." *IEEE transactions on Information Theory* 50.12 (2004): 3250-3264.

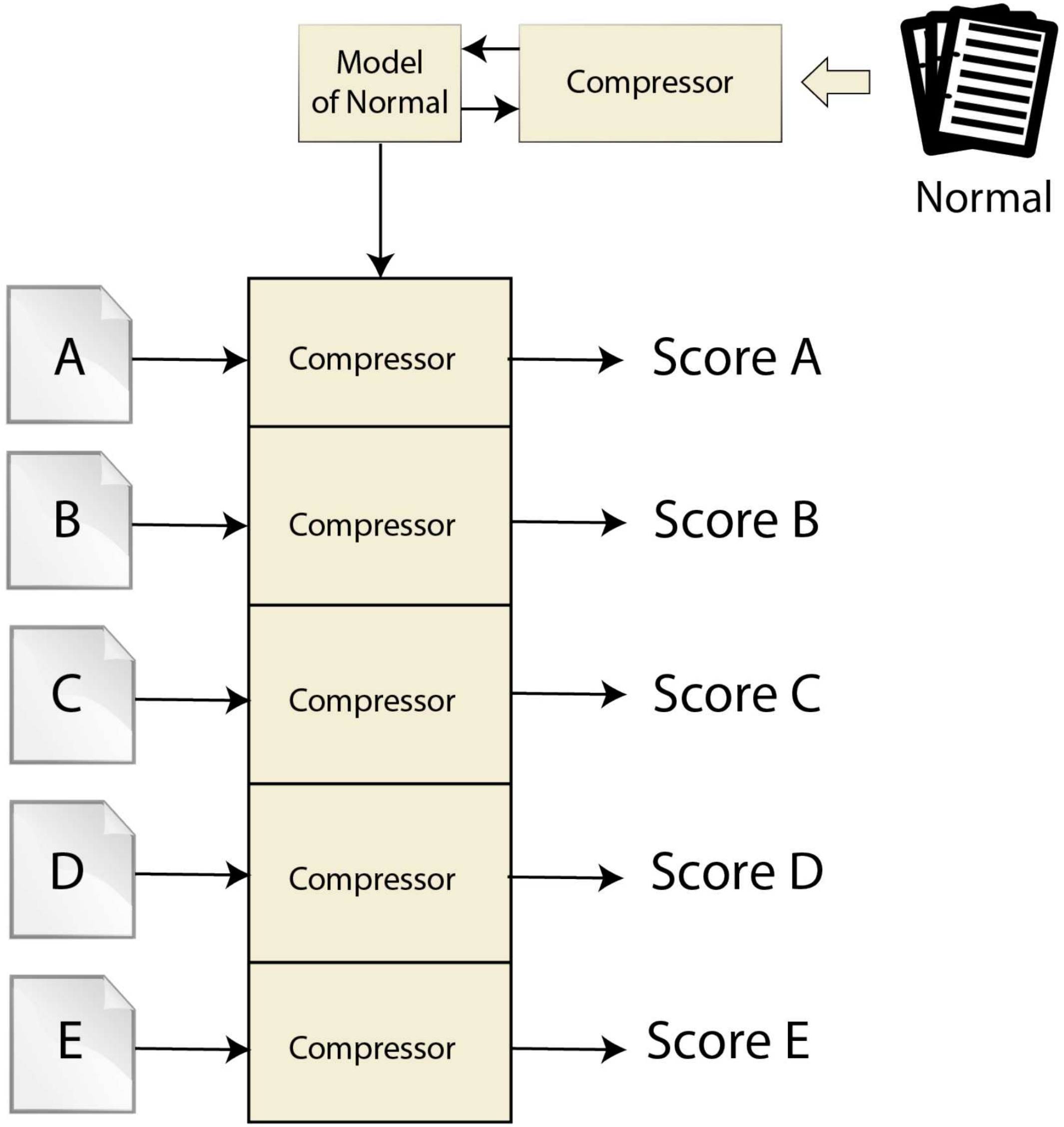
PPM/Arithmetic Coding is straight forward and a flexible implementation lends itself to analytics R&D



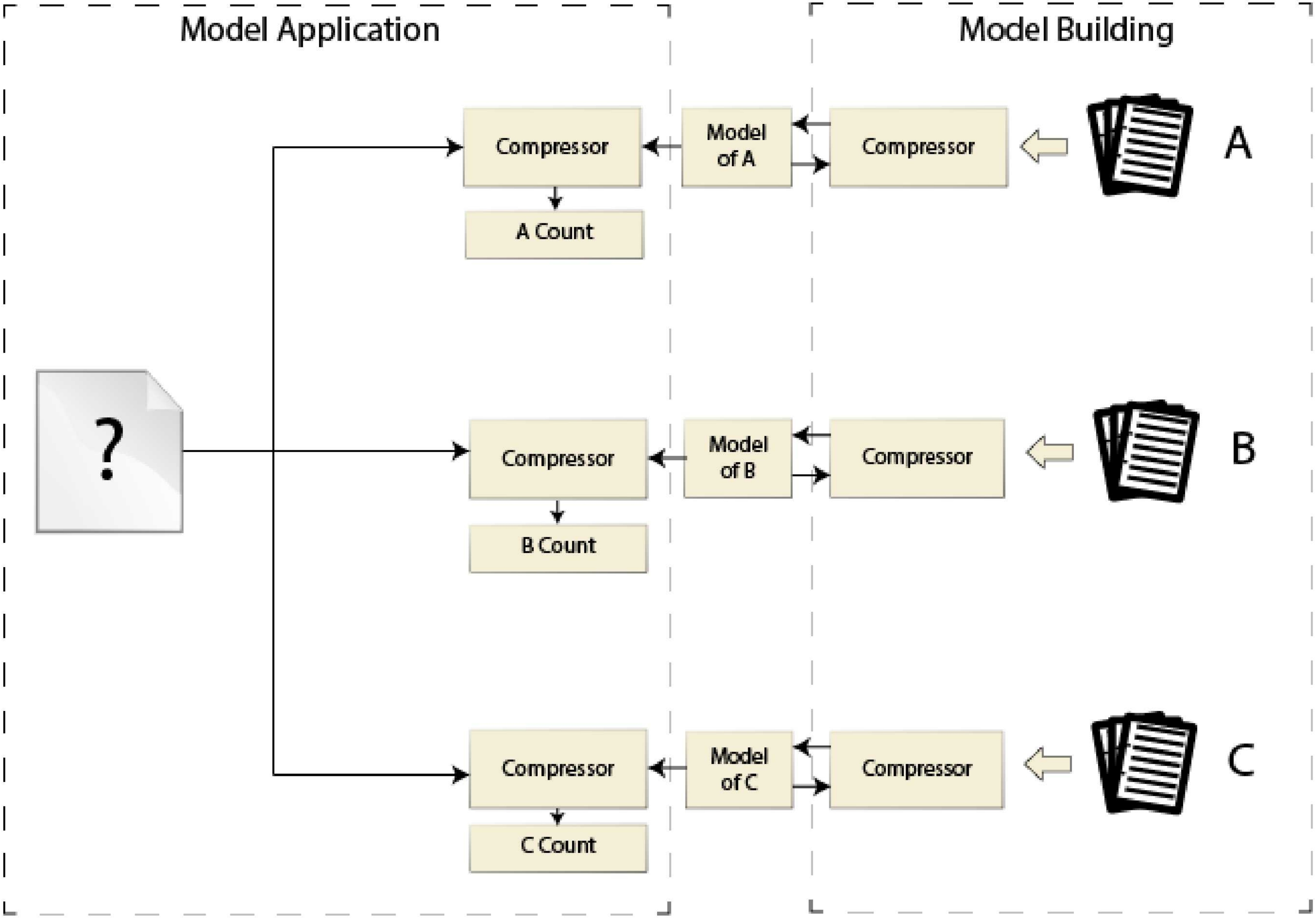
Distances Among Items



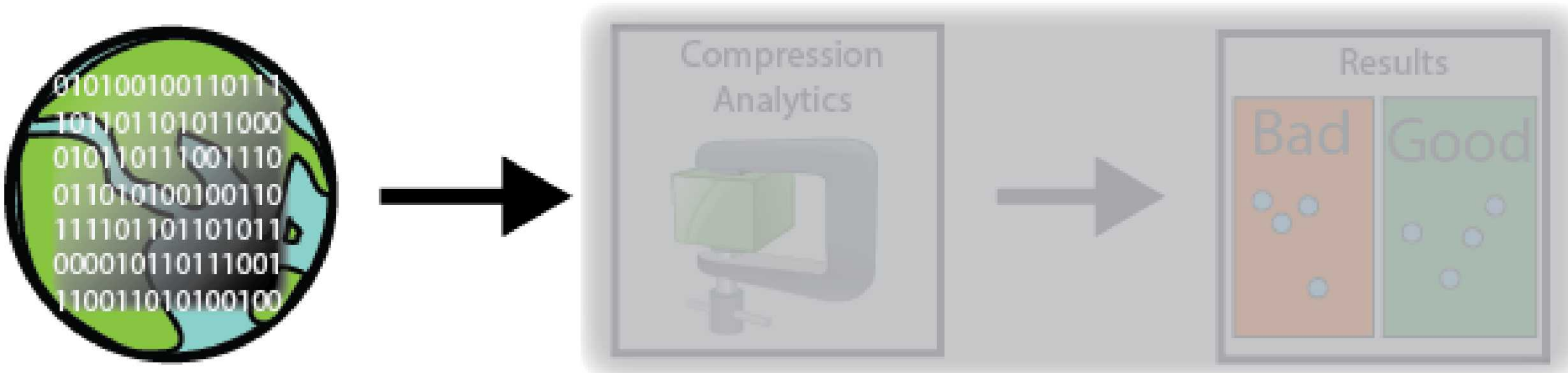
Ordering and Outlier Detection



Categorization



WHERE compression analytics work

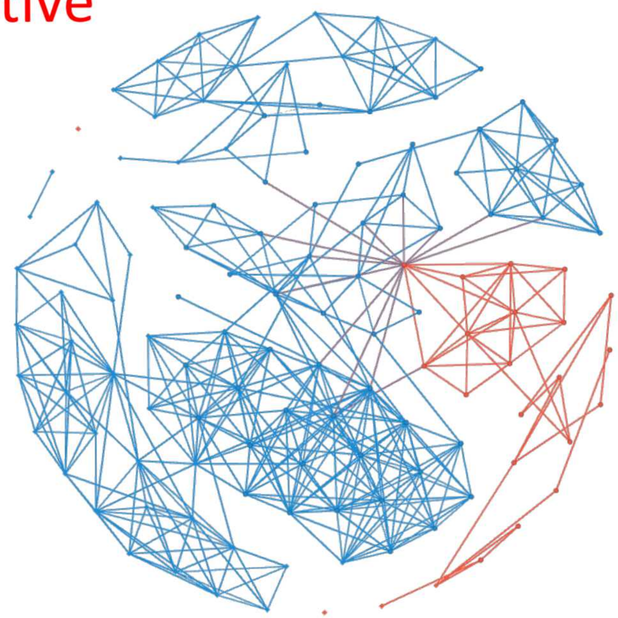


We have successfully applied compression to a variety of different data types and problems.

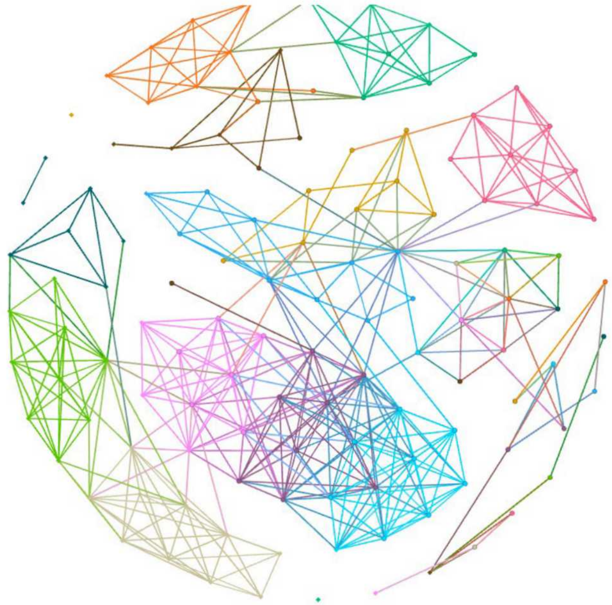
Text

Authorship and Deception Detection

Truthful vs deceptive

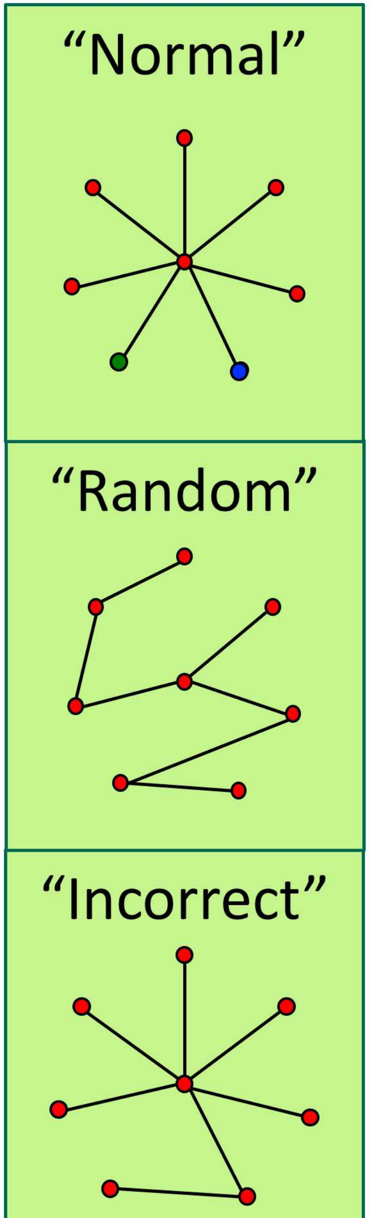


Authorship (12)



$$NCD = \frac{c(xy) - \min[c(x), c(y)]}{\max[c(x), c(y)]}$$

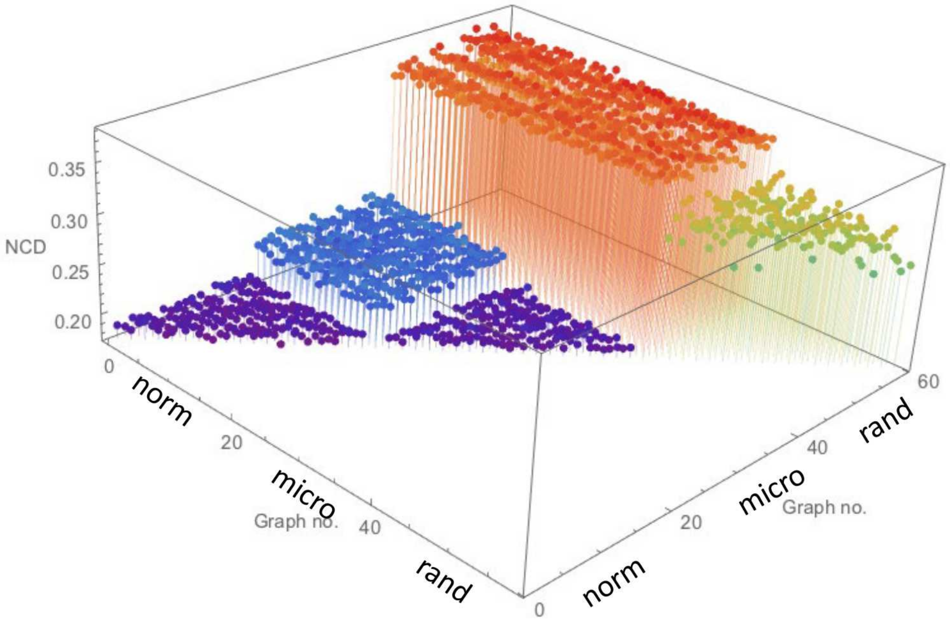
XML Representation of Graphs

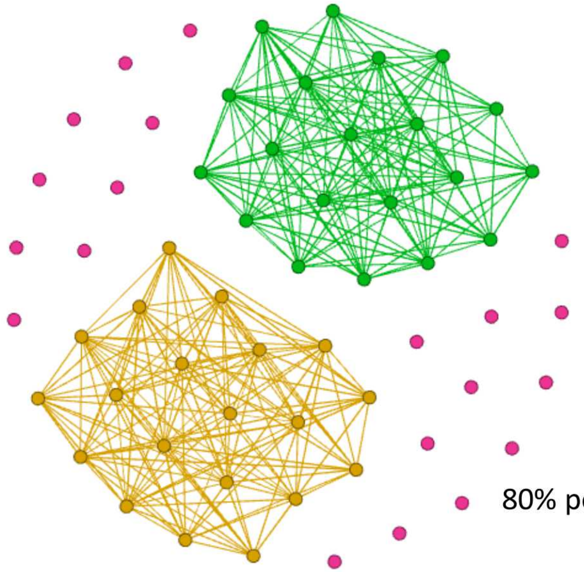


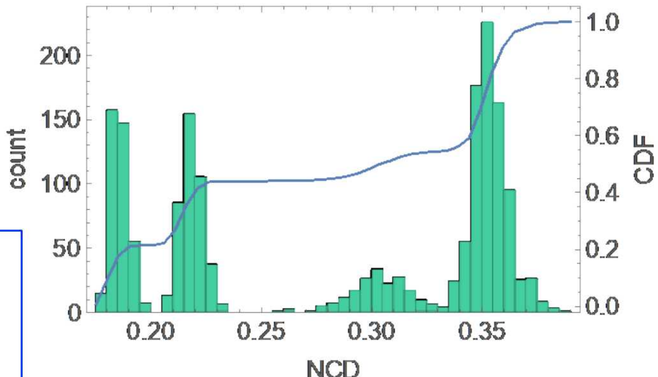

```xml
<gexf xmlns="http://www.gexf.net/1.2draft" version="1.2">
<graph mode="static" defaultedgetype="undirected">
<attributes class="node">
        <attribute id="0" title="label" type="string" />
</attributes>

<nodes>
        <node id="1" label="root: " >
                <attvalues>
                        <attvalue for="0" value="root: toolq"/>
                </attvalues>
        </node>
        <node id="2" label="name: " >
                <attvalues>
                        <attvalue for="0" value="name: asqsv"/>
                </attvalues>
        </node>
        <node id="3" label="age: " >
                <attvalues>
                        <attvalue for="0" value="age: ahglw"/>
                </attvalues>
        </node>
        <node id="4" label="gender: " >
                <attvalues>
                        <attvalue for="0" value="gender: bjwzt"/>
                </attvalues>
        </node>
        <node id="5" label="education: " >
                <attvalues>
                        <attvalue for="0" value="education: zsool"/>
                </attvalues>
        </node>
</nodes>



<edges>
        <edge id="6" source="3" target="1" type="undirected" label="1.0" weight="1.0" />
        <edge id="7" source="2" target="1" type="undirected" label="1.0" weight="1.0" />
        <edge id="8" source="4" target="1" type="undirected" label="1.0" weight="1.0" />
        <edge id="9" source="5" target="4" type="undirected" label="1.0" weight="1.0" />
</edges>
</graph>
</gexf>
```

XML Representation of Graphs

- $n = 10$
- $k = 10$
- A = {`a`,`b`,`c`,`d`,`e`}

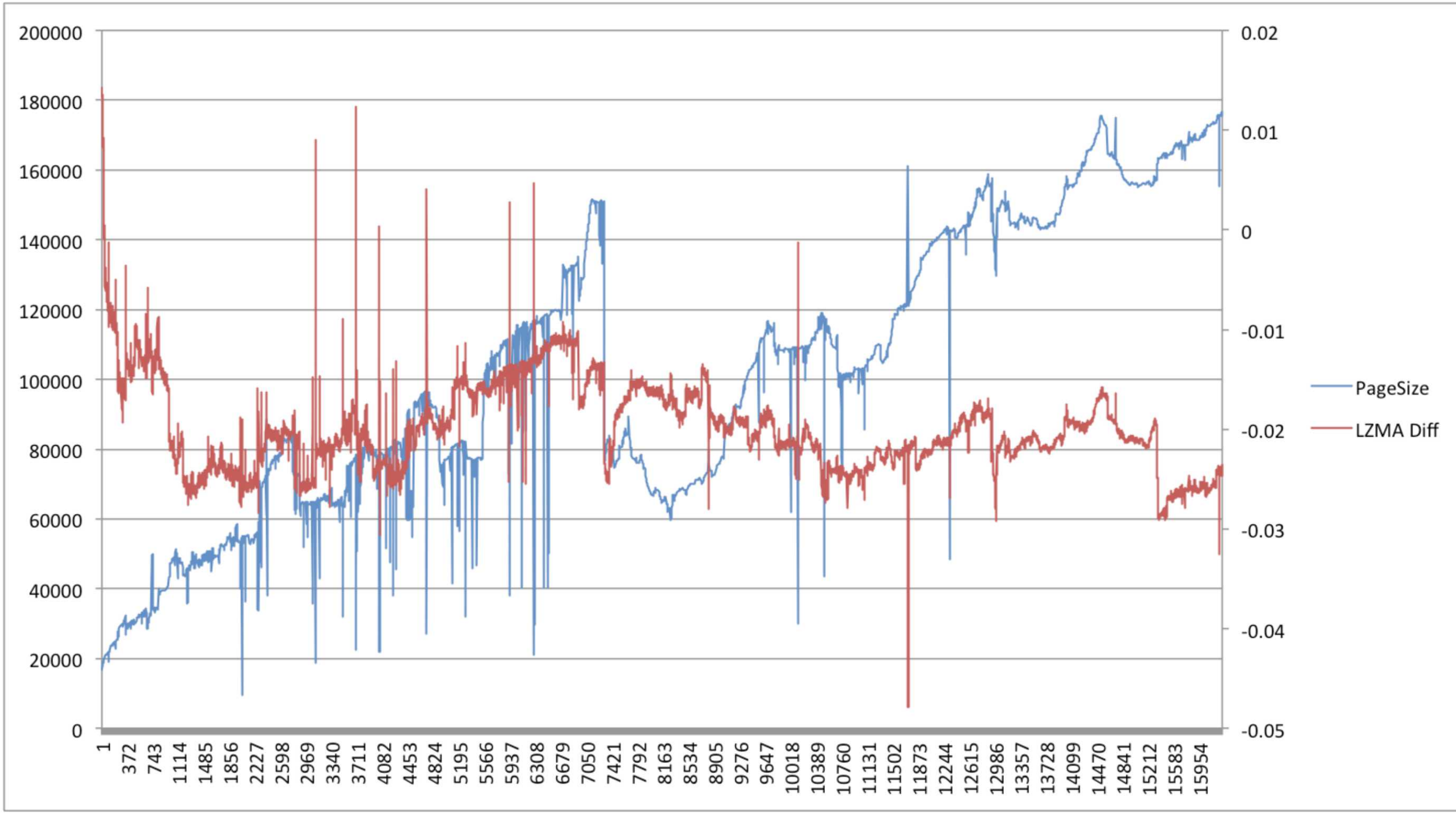






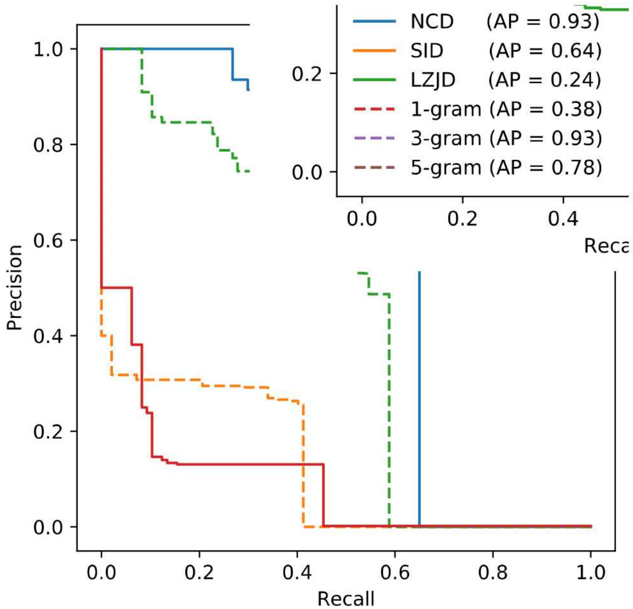
The random string, with k = 10 introduces variance to the NCD scores, but we can still identify a cutoff in the distribution of high similarity NCD scores

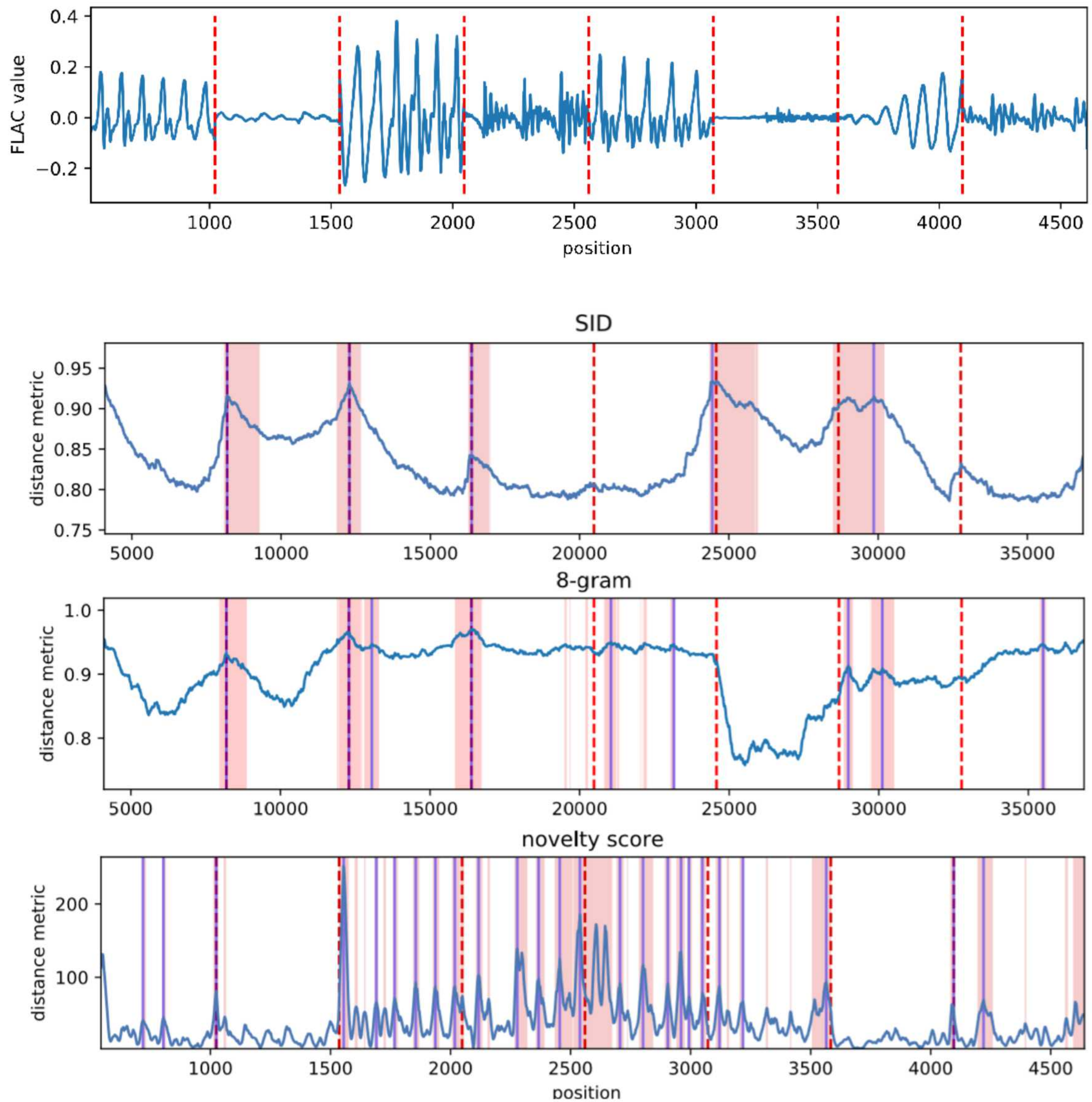
Elvis


200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
1
372
743
1114
1485
1856
2227
2598
2969
3340
3711
4082
4453
4824
5195
5566
5937
6308
6679
7050
7421
7792
8163
8534
8905
9276
9647
10018
10389
10760
11131
11502
11873
12244
12615
12986
13357
13728
14099
14470
14841
15212
15583
15954
PageSize
LZMA Diff

Binary

Voice

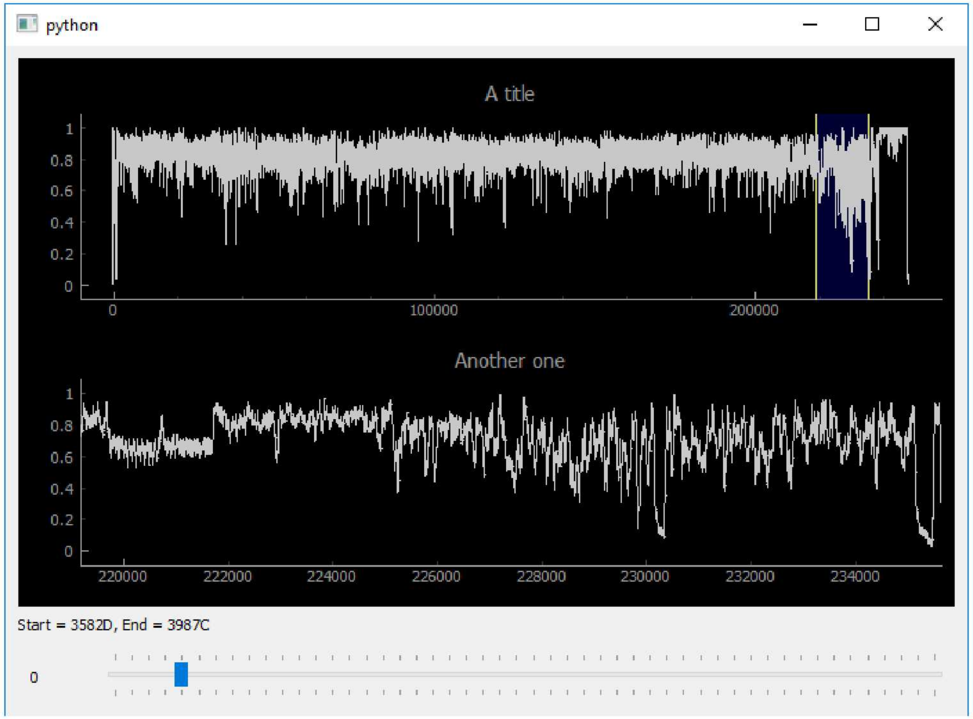

NCD (AP = 0.93)
SID (AP = 0.64)
LZJD (AP = 0.24)
1-gram (AP = 0.38)
3-gram (AP = 0.93)
5-gram (AP = 0.78)
Precision
Recall

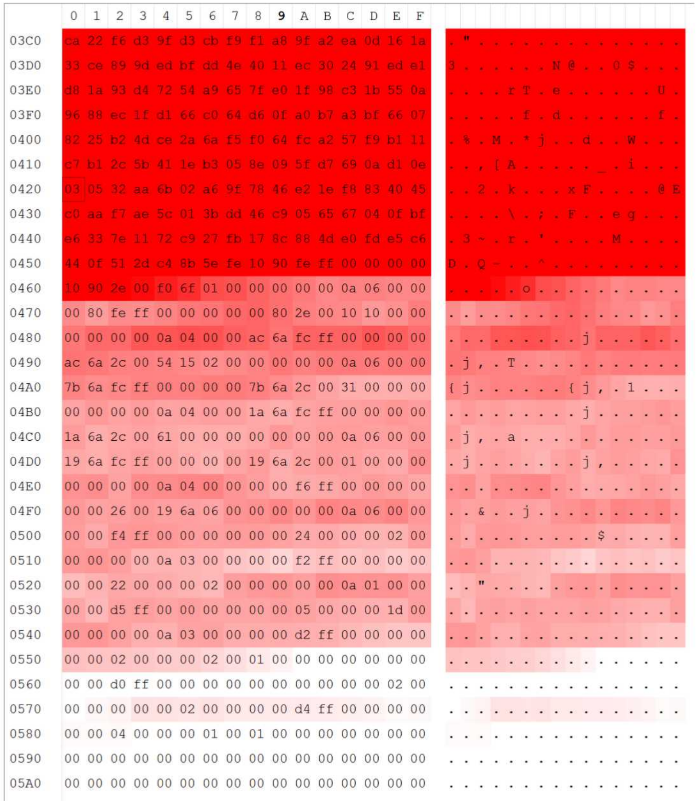

FLAC value
position
SID
distance metric
8-gram
distance metric
novelty score
distance metric
position

Changepoint Detection and Visualization

$$\overbrace{b_{k-w}\, b_{k-w+1} \ldots\ldots\; b_k}^{A}\; \overbrace{b_{k+1} \ldots\ldots\; b_{k+w}}^{B}\; b_{k+w+1}$$

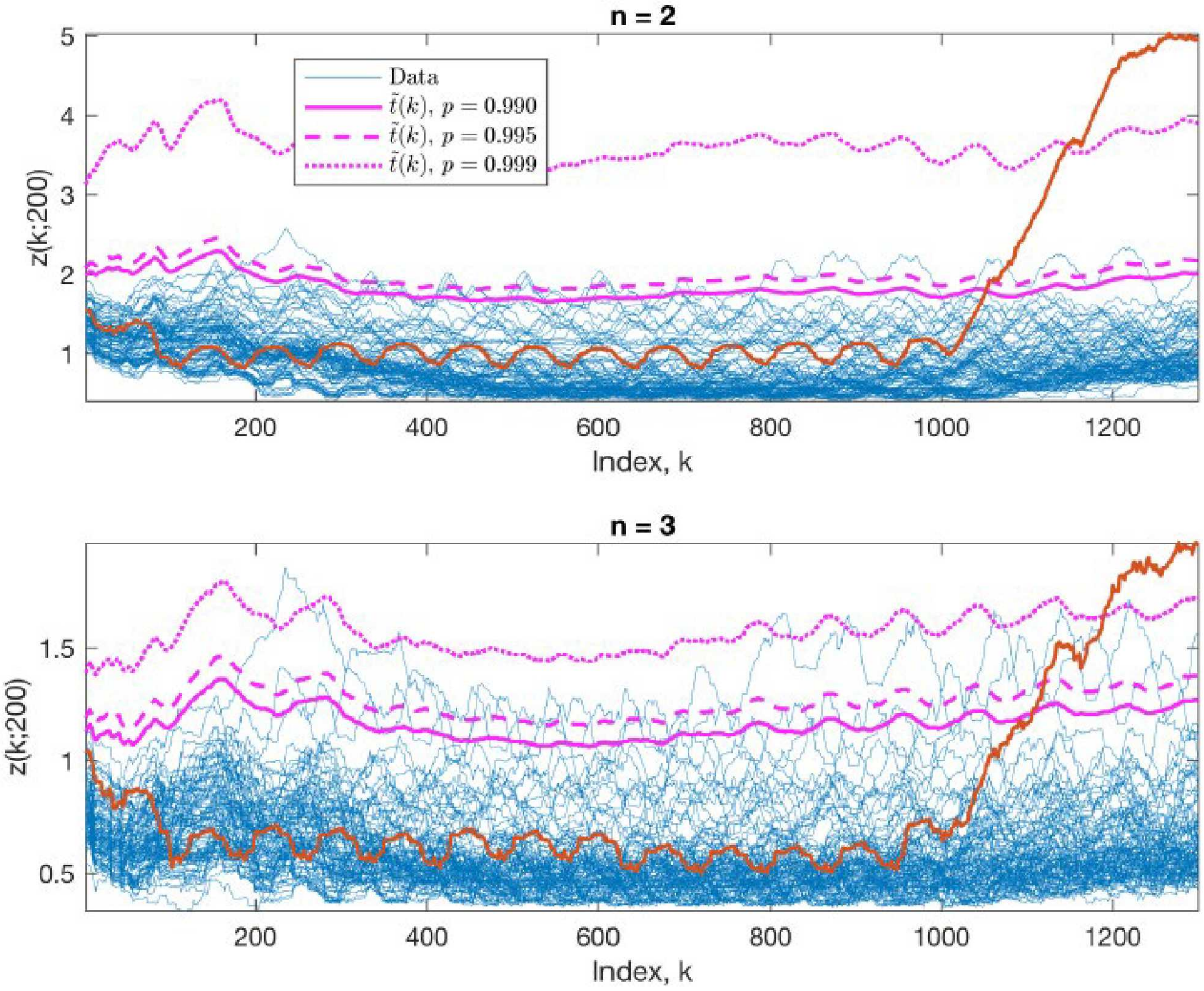
$$\underbrace{b_{k-w+1} \ldots\ldots\; b_k\, b_{k+1}}_{A'}\; \underbrace{\ldots\ldots\; b_{k+w}\, b_{k+w+1}}_{B'}$$



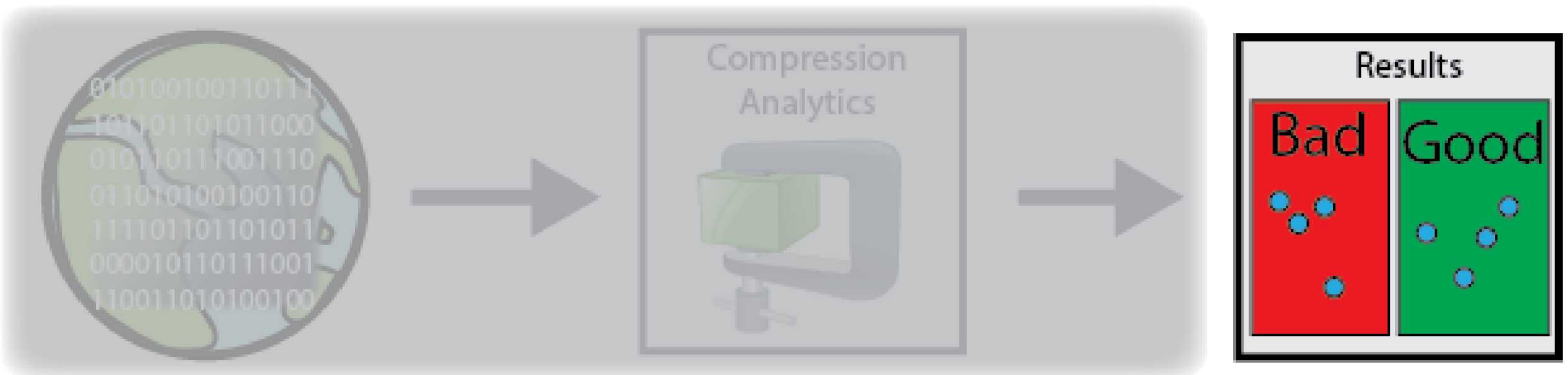


Network Traffic

Analysis of DNS Traffic



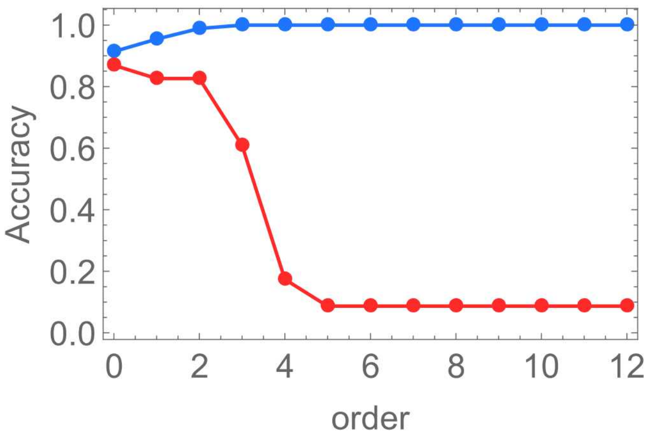
How to make compression analytics work BETTER

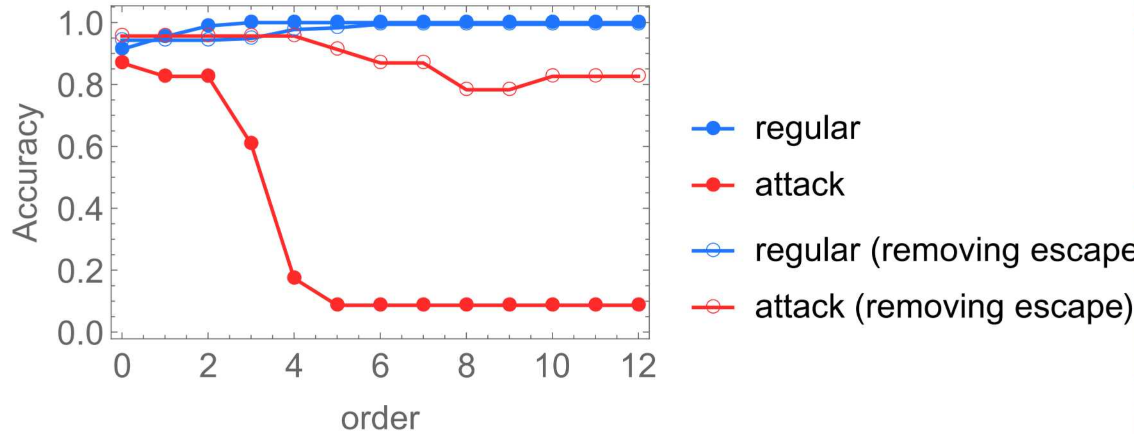


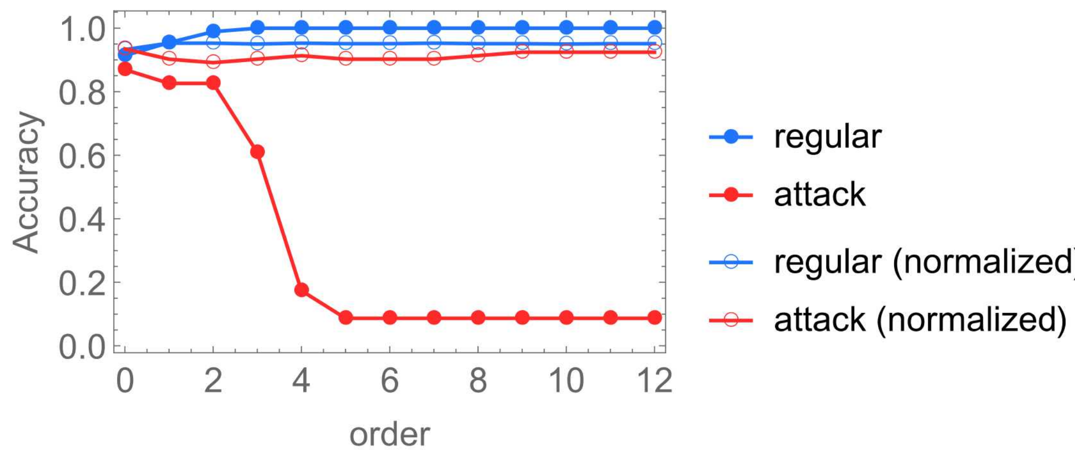
There are interesting remaining problems in advancing the state of the art in this field.

We don’t care about compression, so we can modify the algorithms.

Unbalanced Training Sets







Compression algorithms are slow.

LZJD is fast and has good performance

"This is a question." → {(' '), (' a'), (' q'), **('.')**, ('T'), ('e'), ('h'), ('i'), ('io'), ('is'), ('n'), ('s'), ('st'), ('u')}

"This is another question." → {(' '), (' a'), (' q'), ('T'), ('e'), ('h'), **('he')**, ('i'), ('io'), ('is'), ('n'), **('n.')**, ('o'), **('r')**, ('s'), ('st'), **('t')**, ('u')}

Raff, Edward, and Charles Nicholas. "An Alternative to NCD for Large Sequences, Lempel-Ziv Jaccard Distance." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

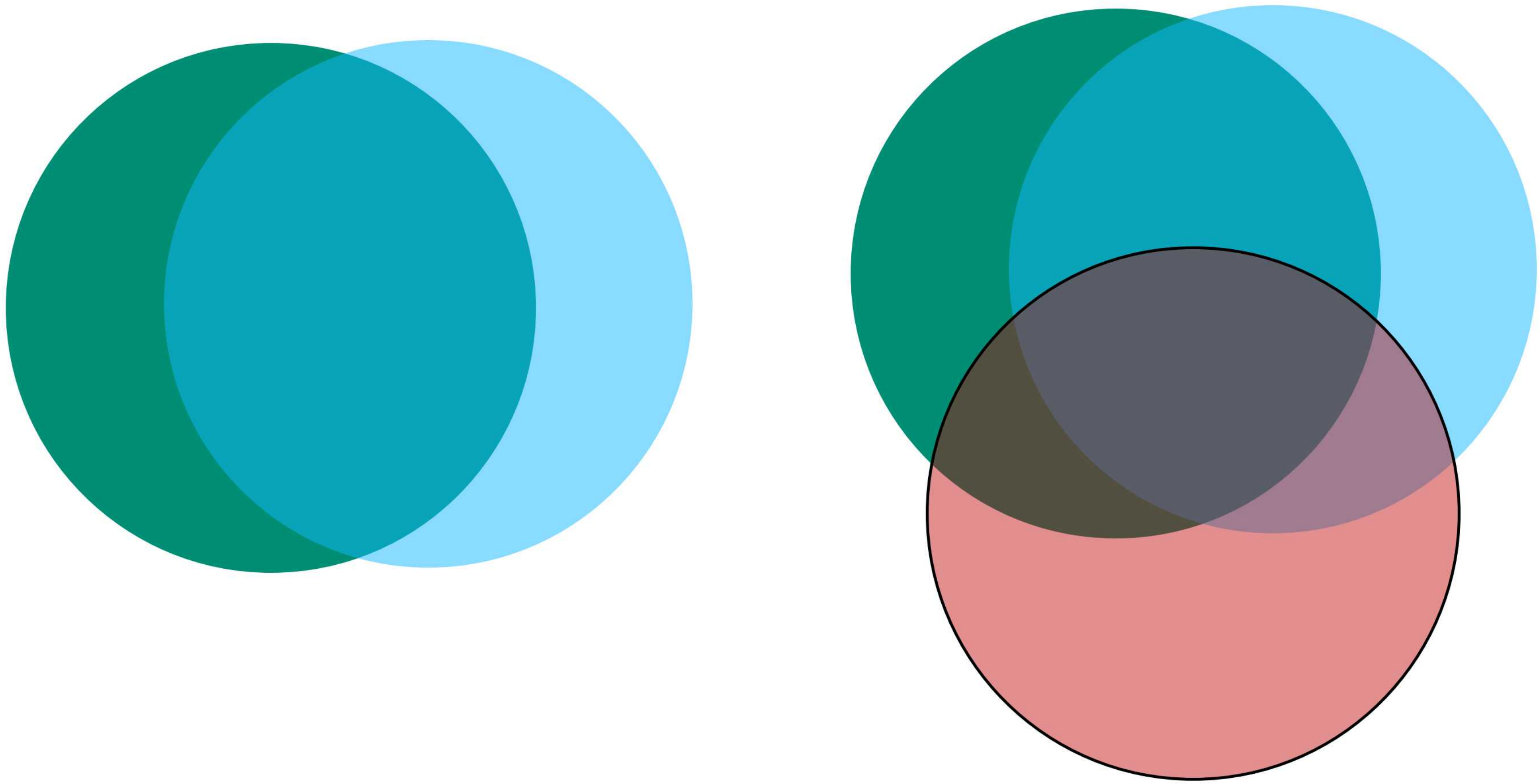
Streaming LZJD is faster when computing sliding windows.

“This is another question.”

[('T'), ('h'), ('i'), ('s'), (' '), ('i'), ('is'),
(' '), (' a'), ('n'), ('o'), ('t'), ('h'),
('he'), ('r'), (' '), (' q'), ('u'), ('e'),
('s'), ('st'), ('i'), ('io'), ('n’)]

You don’t have to specify the features, but you can’t specify the features.

NCD with Controls



Non Sequential Analysis

This is a question.

<p>This is HTML</p>

Contact Information and Relevant Publications and Reports

Travis Bauer (tlbauer@sandia.gov)

Publications

- McNamara, L., Bauer, T., Haass, M., and Matzen, L., "Silver Ticket? Exploring Entropy Metrics for Visual Analytics." BELIV 2016, Oct 24, 2016, Baltimore.

- Brounstein, Tom Rego, et al. *Stylometric and Temporal Techniques for Social Media Account Resolution*. No. SAND2017-2965C. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2017.

- Ting, C., Fisher, A., Bauer, T., "Compression based algorithms for deception detection." SocInfo2017, Sep 13-15, 2017, Oxford, UK

- Ting, C., Field, R., Fisher, A., Bauer, T., “Compression Analytics for Classification and Anomaly Detection within Network Communication.” IEEE Transactions on Information Forensics & Security Volume 14, Issue 5, May 2019

- Ting, C., Field, R., Quach, T., Bauer, T., “Generalized Boundary Detection in Streaming Data Using Compression-Based Analytics. ICASSP 2019, accepted