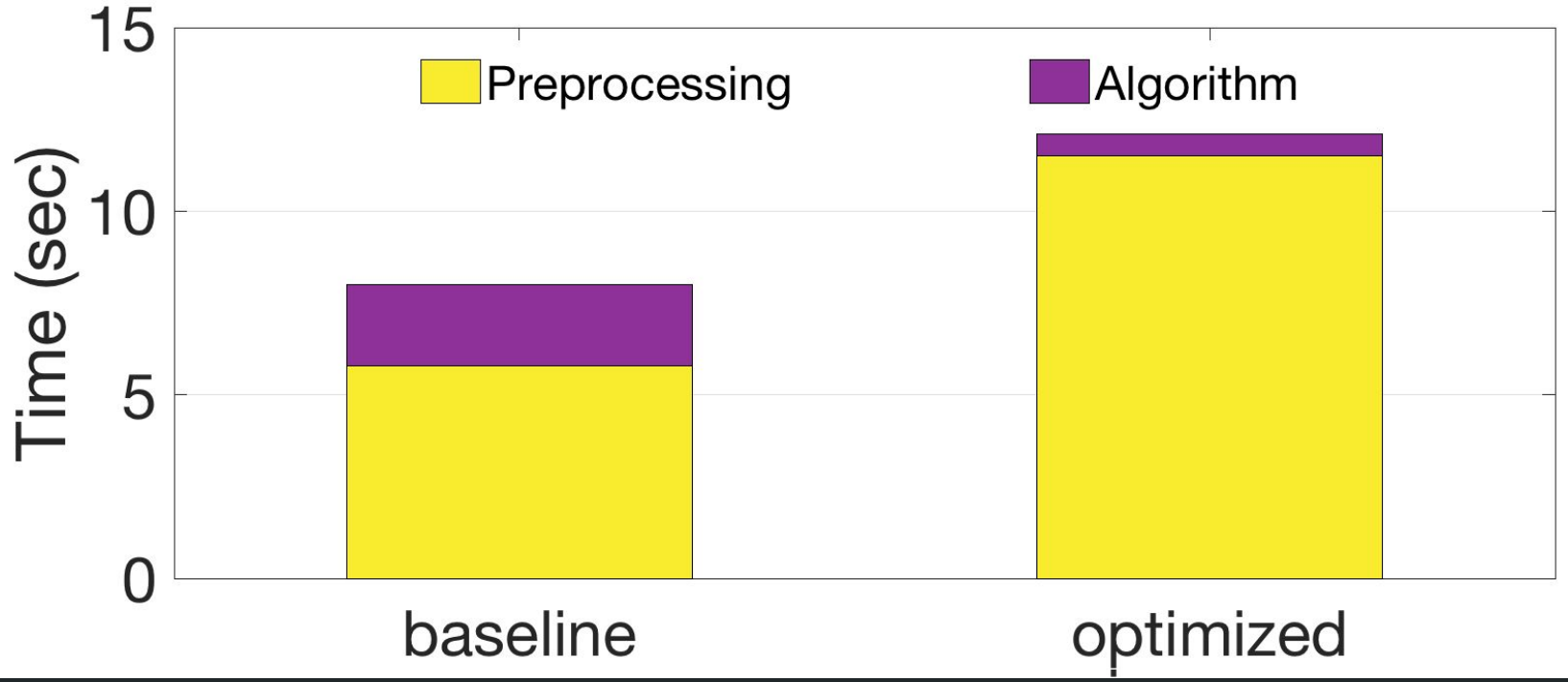
DIBS: A Data Integration Benchmark Suite

AM Cabrera, CJ Faber, K Cepeda, R Derber, C Epstein, J Zheng,
RK Cytron, and RD Chamberlain
Department of Computer Science and Engineering
Washington University in St. Louis, MO, USA

The Preprocessing Pain Point

BFS on Twitter Data



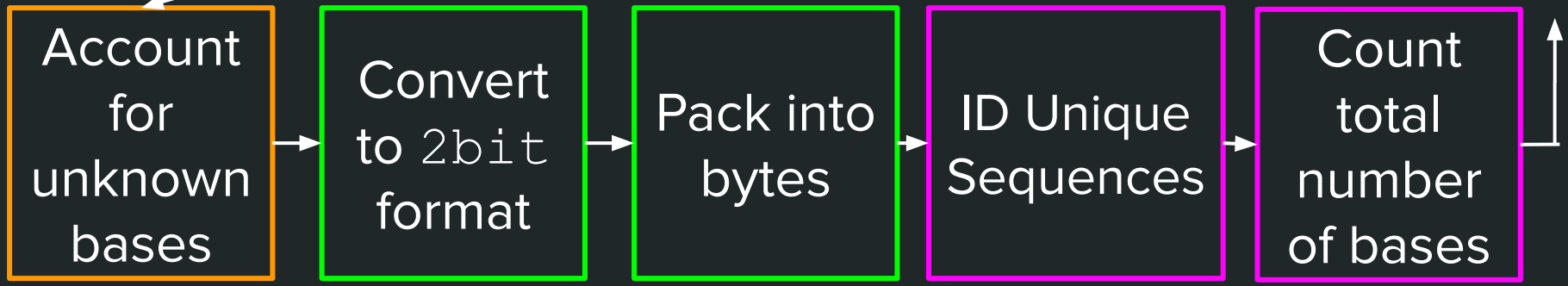
Data Integration

Parsing/Cleansing

Transformation

Aggregation

```
>Some Sequence of Interest
...
agcaagacttcatctcaaaaaaaaaaaaaaaGCTGCANATTTattattat
tattattagtttatttatttatttttttgagacagagtctcgttctgtcg
cccaggctggagtgcggtggcgtgatcttggctcattgcaacctccacct
cccgggttcaagtgattctcctgcctcagcctcccgagtagctgggacta
caggcgtatgccaccatgcctggctaatttttttgtacttttagtagagac
Agagtttcacggtgttagccaggctggtcttgatctcctgacctcgtgat
...
```



Our Contribution

Develop a data integration benchmarking suite (DIBS) through creating a comprehensive set of data integration tasks

Develop characterization to provide insights regarding research across the hardware and software stack.

Make publicly available to allow researchers to develop solutions and evaluate their performance

	Data Integration Tasks		
Domain	**Parsing/Cleansing**	**Transformation**	**Aggregation**
Computational Biology		fa → 2bit 2bit → fa	
Image Processing		fits → tiff idx → tiff optdigits → tiff unipen → tiff	
Enterprise		ebcdic → txt fix → float	
Internet of Things		tstcsv → csv gotrackcsv → csv plt → csv	
Graph Processing		edgelist → csr	

	Data Integration Tasks		
Domain	**Parsing/Cleansing**	**Transformation**	**Aggregation**
Computational Biology	Separate bases Handle unclear bases	fa → 2bit 2bit → fa	Track total size
Image Processing	Parse FITS Tags	fits → tiff idx → tiff optdigits → tiff unipen → tiff	Pixel stats Pixel adjustment Histogram
Enterprise	Adjust non-ASCII characters	ebcdic → txt fix → float	Count elements
Internet of Things	Tokenize input	tstcsv → csv gotrackcsv → csv plt → csv	Calculate file size
Graph Processing	Parse edge list	edgelist → csr	Count nodes/edges Compute degree

Application Characterization

Characteristic Dimensions

- Locality
- Branch Entropy
- Instruction Mix

Locality

Spatial Locality

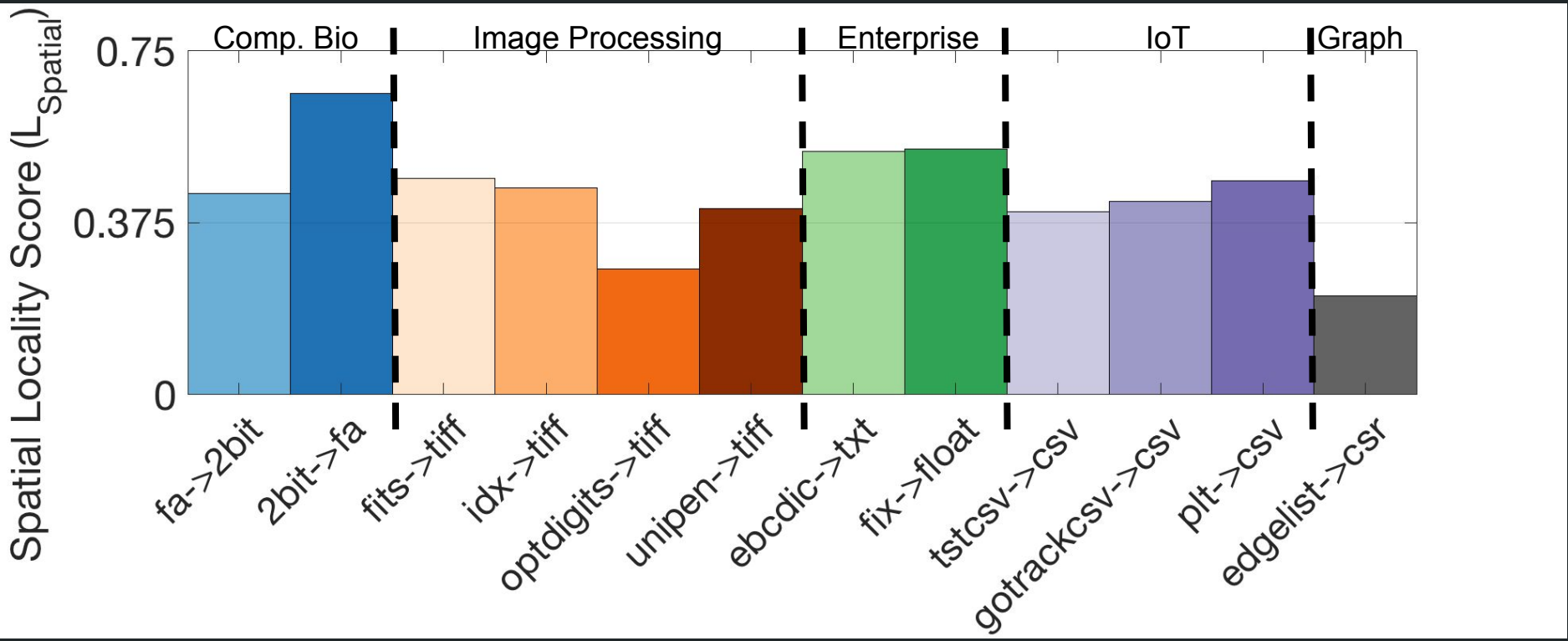
Temporal Locality

$$L_{Spatial} = \sum_{i=1}^{\infty} \frac{stride_i}{i}$$

$$L_{Temporal} = \frac{\sum_{i=0}^{\log_2(N)-1} \left[(reuse_{2^{i+1}} - reuse_{2^i}) \times (\log_2(N) - i) \right]}{\log_2 N}$$

Weinberg et al. SC ‘08

Spatial Locality

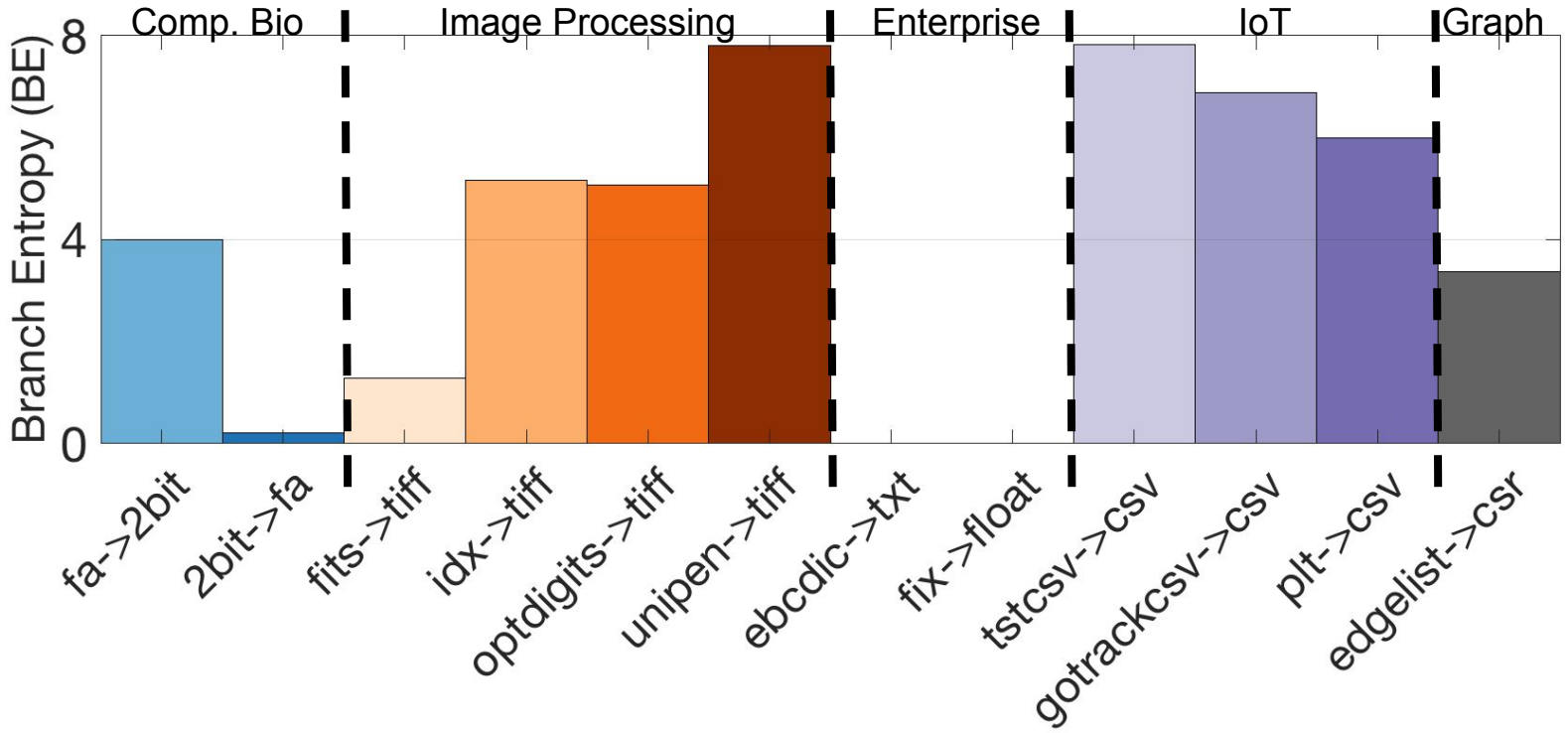


Branch Entropy

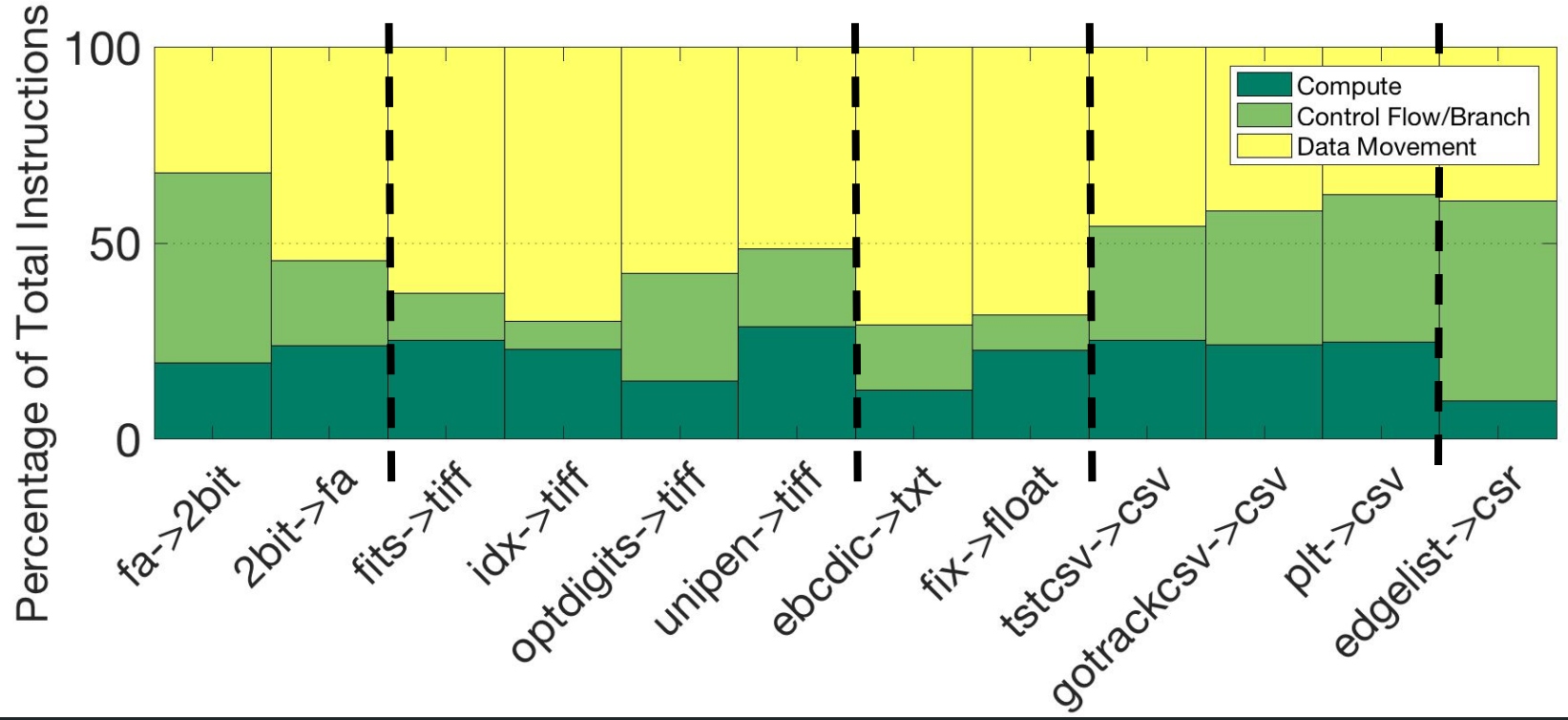
$$BE = -\sum_{i} p(E_i) \log_2 p(E_i)$$

Yokota et al. ARCS ‘08

Branch Entropy



x86-64 Instruction Mix



Conclusion

Data Integration Benchmarking Suite

Quantitative Characterization

Consistency in Locality

Control Flow Regularity

Prevalence of Data Movement

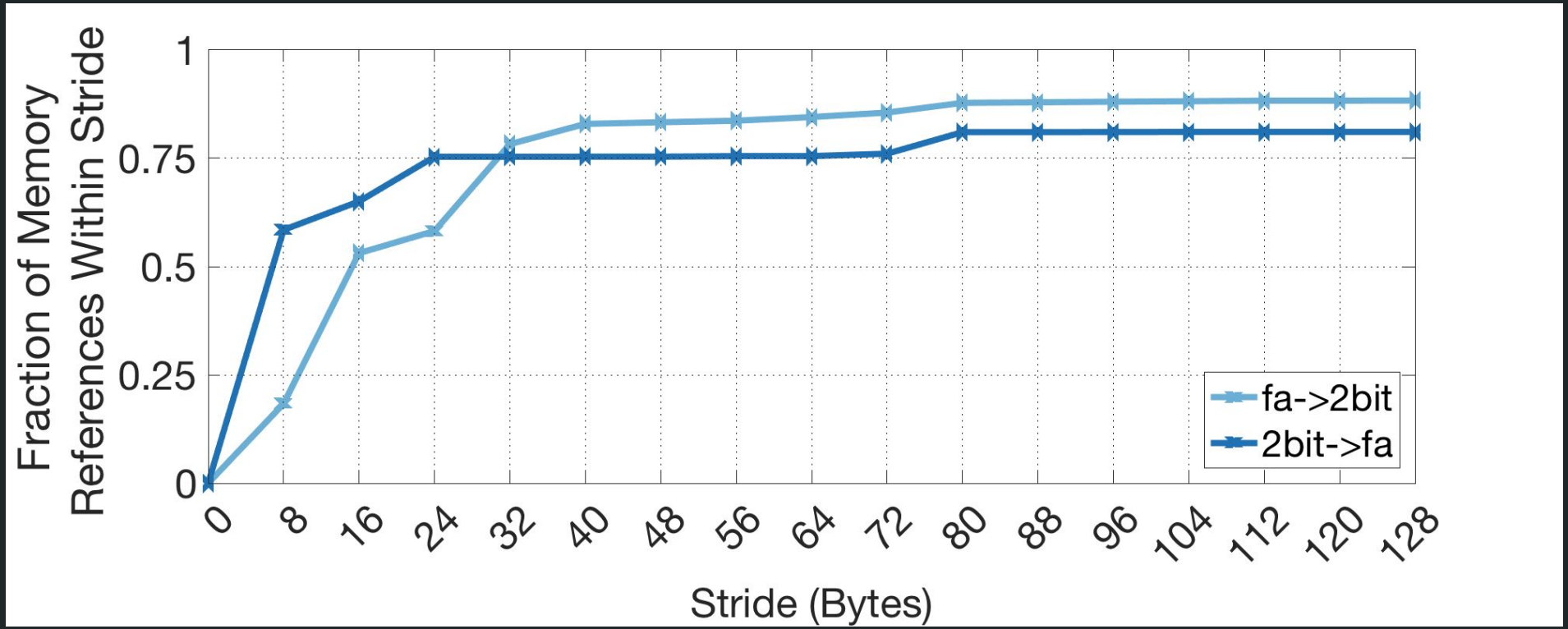
Questions?

DIBS + Data: https://openscholarship.wustl.edu/data/9/

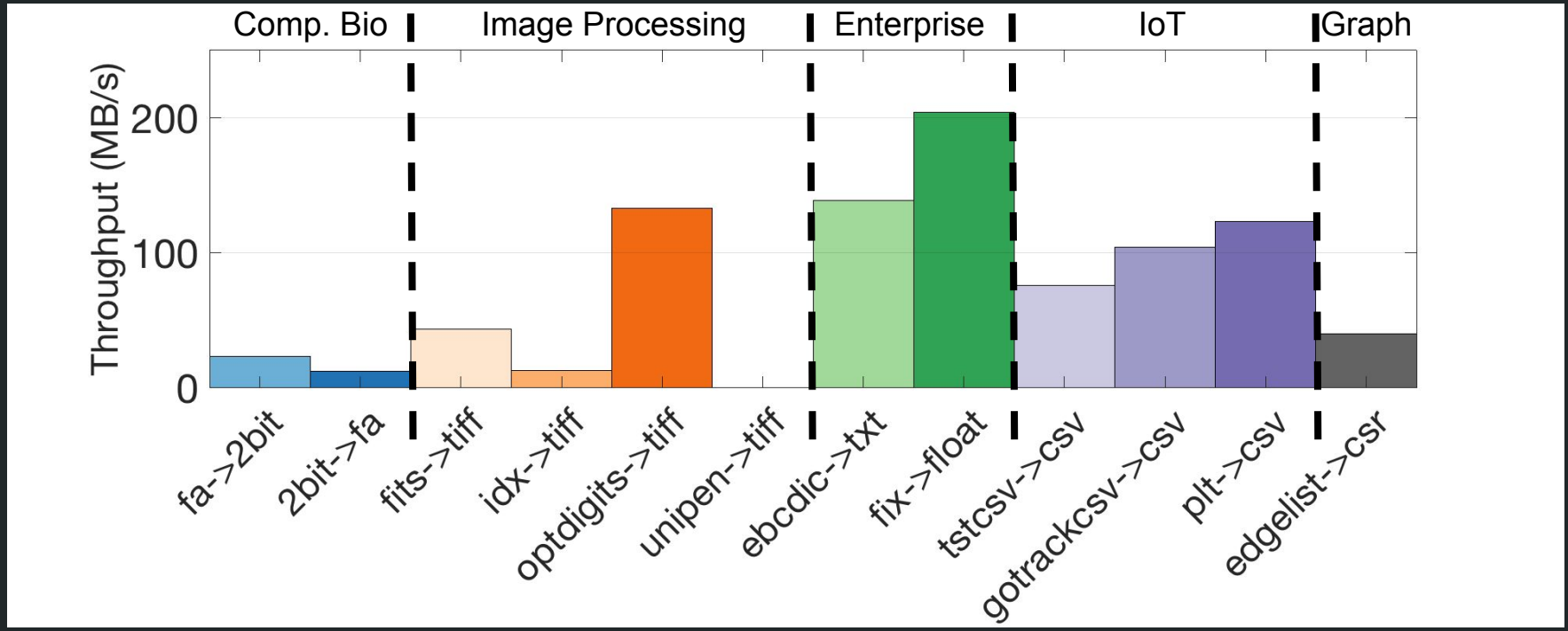
My Page: https://sites.wustl.edu/acabrera/


WASHINGTON UNIVERSITY IN ST. LOUIS
1853

Cumulative Sum for Computational Biology Applications



Throughput


Comp. Bio
Image Processing
Enterprise
IoT
Graph
Throughput (MB/s)
200
100
0
fa->2bit
2bit->fa
fits->tiff
idx->tiff
optdigits->tiff
unipen->tiff
ebcdic->txt
fix->float
tstcsv->csv
gotrackcsv->csv
plt->csv
edgelist->csr

Temporal Locality