# Hashing Algorithms for Large-Scale Learning

- **Ping Li**, Dept. of Statistical Science, Cornell University
- **Anshumali Shrivastava**, Dept. of Computer Science, Cornell University
- **Joshua Moore**, Dept. of Computer Science, Cornell University
- **Christian König**, Microsoft Research, Microsoft Corporation

## Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Examples ($n$)</th>
<th># Dims ($D$)</th>
<th>Avg # Nonzeros</th>
<th>Train / Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Webspam (24 GB)</td>
<td>350,000</td>
<td>16,609,143</td>
<td>3728</td>
<td>80% / 20%</td>
</tr>
<tr>
<td>Rcv1 (200 GB)</td>
<td>781,265</td>
<td>1,010,017,424</td>
<td>12062</td>
<td>50% / 50%</td>
</tr>
</tbody>
</table>
Experiments on Webspam Data (24GB, 16M features)

- Solid: $b$-bit hashing. Dashed (red) the original data.
- Using $b = 8$ and $k \geq 200$ permutations (about 70 MB only) achieves similar test accuracies as using the original data.
Comparisons with Random Projections and Vowpal Wabbit (VW)

The theoretical variance of $b$-bit minwise hashing is substantially lower than random projections and Vowpal Wabbit (VW). For webspam, $b = 8$ and $k = 200$ achieved similar accuracies as VW with $k > 2^{16}$ hashing bins.

Note that $2^8 \times 200 < 2^{16}$. Therefore, $b$-bit minwise hashing simultaneously solved two problems: (i) Data storage; (ii) Dimension reduction.
Experiments on Expanded RCV1 Data (200GB, 1 billion features)

Test accuracy using linear SVM (Can not train the original data)

Comparisons with Vowpal Wabbit (VW)